Journal of Computer-Aided Molecular Design, 6 (1992) 253–272
ESCOM

J-CAMD 156

PLS modelling of structure–activity relationships of catechol *O*-methyltransferase inhibitors

Timo Lotta, Jyrki Taskinen*, Reijo Bäckström and Erkki Nissinen

Orion Research Center, Orion Corporation, Orion-Farmos Pharmaceuticals, P.O. Box 65, SF-02101 Espoo, Finland

Received 25 September 1991
Accepted 30 December 1991

Key words: QSAR; Computational chemistry; PLS; Multivariate; COMT; Catechols

SUMMARY

Quantitative structure–activity analysis was carried out for in vitro inhibition of rat brain soluble catechol *O*-methyltransferase by a series (N=99) of 1,5-substituted-3,4-dihydroxybenzenes using computational chemistry and multivariate PLS modelling of data sets. The molecular structural descriptors (N=19) associated with the electronics of the catecholic ring and sizes of substituents were derived theoretically. For the whole set of molecules two separate PLS models have to be used. A PLS model with two significant (crossvalidated) model dimensions describing 82.2% of the variance in inhibition activity data was capable of predicting all molecules except those having the largest R_1 substituent or having a large R_5 substituent compared to the NO_2 group. The other PLS model with three significant (crossvalidated) model dimensions described 83.3% of the variance in inhibition activity data. This model could not handle compounds having a small R_5 substituent, compared to the NO_2 group, or the largest R_1 substituent. The predictive capability of these PLS models was good. The models reveal that inhibition activity is nonlinearly related to the size of the R_5 substituent. The analysis of the PLS models also shows that the binding affinity is greatly dependent on the electronic nature of both R_1 and R_5 substituents. The electron-withdrawing nature of the substituents enhances inhibition activity. In addition, the size of the R_1 substituent and its lipophilicity are important in the binding of inhibitors. The size of the R_1 substituent has an upper limit. On the other hand, ionized R_1 substituents decrease inhibition activity.

INTRODUCTION

Catechol *O*-methyltransferase (COMT; EC 2.1.1.6) is an enzyme that catalyzes the transfer of a methyl group from *S*-adenosyl-L-methionine (SAM) to one of the phenolic hydroxyl groups of catechol or substituted catechols [1]. COMT is widely distributed both in the peripheral and in the central nervous system [2]. COMT has an important role in extraneuronal inactivation of, for in-

* To whom correspondence should be addressed.

stance, drugs with catechol-like structure such as L-DOPA and a large number of other catechol substrates [1–4]. The occurrence of two distinct isoforms of COMT has been demonstrated; one is soluble COMT (S-COMT) and the other is membrane-bound COMT (MB-COMT) [5–10]. The present study involves only experiments with the soluble COMT enzyme.

The investigation of the physiological role of COMT as a dopamine and noradrenaline metabolizing enzyme in the central nervous system was for a long time hindered by the lack of effective COMT inhibitors. However, the development of a family of very potent COMT inhibitors during the last few years has now led to the use of COMT inhibitors, in combination with L-DOPA, in clinical studies of Parkinson's disease [11–13].

A quantitative structure–activity relationship analysis (QSAR) has been carried out for in vitro inhibition of rat liver S-COMT by twenty-three 1,5-substituted-3,4-dihydroxybenzenes [14]. Linear regression models showed high correlation of inhibitory activity with both empirical electronic substituent parameters and quantum-chemical descriptors [14].

Traditionally in QSAR, multiple linear regression (MLR) and related models are used for data analysis. The use of regression analysis in QSAR studies has some well-known problems. For instance, ideally, a statistical experimental design should be used. The number of descriptors included in the model must be kept small compared with the number of compounds [15]. In the present work QSAR analysis was carried out for in vitro inhibition of rat brain S-COMT by a series (N = 99) of 1,5-substituted-3,4-dihydroxybenzenes using the PLS data analytic method (PLS; partial least squares with latent variables). The general structure of the S-COMT inhibitors is shown in Fig. 1. These compounds accumulated during the synthetic project and do not show a balanced coverage of substituent space. Furthermore, substituent constants are not available for most compounds. However, useful information may be extracted from this kind of material if appropriate statistical methods [15–19] and computed structural parameters are used.

MATERIALS AND METHODS

Compounds

Compounds were synthesized at the Synthetic Department of Orion Research Center, Orion-Farmos Pharmaceuticals (Espoo, Finland). The chemical structures of the compounds are shown in Table 1 and Fig. 1. The training sets for PLS models 1, 3 and 4 were: model 1 (the training set: 4, 8, 10, 12, 19, 21, 22, 23, 25, 29, 30, 42, 45, 47, 49, 54, 57, 66, 74, 75, 77, 78, 79, 82, 86, 89, 90, 91, 92, 93, 98 and 99), model 3 (the training set: 4, 8, 10, 12, 19, 22, 25, 29, 30, 42, 45, 47, 54, 57, 66, 74, 75, 77, 78, 79, 82, 84, 85, 90, 93, 98 and 99), and model 4 (the training set: 4, 8, 10, 12, 16, 19, 22, 25, 29, 30, 42, 45, 47, 54, 57, 66, 74, 75, 77, 83, 85, 86, 87, 89, 90, 91 and 92). The numbering of compounds is presented in Table 1.

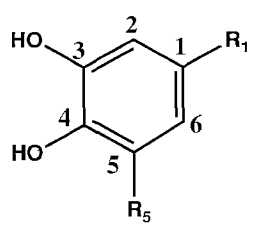

Fig. 1. The chemical structure of S-COMT inhibitors.

TABLE 1
SUBSTITUENTS OF COMPOUNDS USED IN THE PRINCIPAL COMPONENT AND PLS ANALYSIS

Cmpd.	R_1	R_5	Cmpd.	R_1	R_5
1	$(CH_2)_4COOH$	NO_2	51	$CHC(CN)CONHCH(CH_3)_2$	NO_2
2	CHCHOOH	NO_2	52	CHC(CN)CO-(4-methyl-1-piperazinyl)	NO_2
3	CHCHCO-phenyl	NO_2	53	$CHC(CN)CON(CH_2CH_3)_2$	NO_2
4	$(CH_2)_4CONH$-(1-adamantyl)	NO_2	54	$CHC(CN)COOCH_2C(CH_3)_3$	NO_2
5	$(CH_2)_4CONHCH(CH_3)_2$	NO_2	55	$CHC(CN)CONH(CH_2)_4OH$	NO_2
6	$CHC(CH_3)CH(OH)CH_3$	NO_2	56	$CONH(CH_2)_3OH$	NO_2
7	$CHC(CH_3)COCH_3$	NO_2	57	COOH	NO_2
8	$CHC(CN)_2$	NO_2	58	$CHC(COCH_3)CO$-phenyl	NO_2
9	$(CH_2)_4CON(CH_3)CH_2CCH$	NO_2	59	CO-(2-pyridyl)	NO_2
10	CHCHCO-(3,4,5-trimethoxyphenyl)	NO_2	60	CHCNHCSNHCO	NO_2
11	CONH-(1-adamantyl)	NO_2	61	$CHC(CN)CONHC(CH_3)_3$	NO_2
12	$CHCHNO_2$	NO_2	62	CHC(CN)CO-(1-piperidyl)	NO_2
13	$CH_2CH(CN)_2$	NO_2	63	CHCCONHCSS	NO_2
14	CHCH-(4-pyridyl)	NO_2	64	$CHC(CN)CON(C_3H_7)_2$	NO_2
15	CHCH-(4-quinolinyl)	NO_2	65	$CHC(COCH_3)CONHCH_2CH_2OH$	NO_2
16	$COCH_3$	NO_2	66	CHCCONHCONHCO	NO_2
17	CHO	NO_2	67	$CH(COCH_3)$-(2-pyridyl)	NO_2
18	$CHC(COCH_3)_2$	NO_2	68	CO-(4-pyridyl)	NO_2
19	COCHCH-(4-(dimethylamino)phenyl)	NO_2	69	$CHC(COCH_3)CON(CH_3)_2$	NO_2
20	CHCHCO-(2-hydroxyphenyl)	NO_2	70	$CHC(COCH_3)CON(C_2H_5)_2$	NO_2
21	$(CH_2)_4CO$-(4-benzyl-1-piperazinyl)	NO_2	71	CHC(CN)-(2-pyridyl)	NO_2
22	CN	NO_2	72	CO-(2-fluorophenyl)	NO_2
23	$(CH_2)_4COO(CH_2)_{13}CH_3$	NO_2	73	$CHC(COCH_3)CONH$-cyclohexyl	NO_2
24	CHCHCO-(4-methylphenyl)	NO_2	74	$CHC(COCH_3)CONH$-1-adamantyl	NO_2
25	NO_2	NO_2	75	CHC(CN)CO-(3,4,5-trimethoxyphenyl)	NO_2
26	CO-(4-(cyclohexylcarbonyl)-1-piperidyl)	NO_2	76	$CHC(COCH_2CH_3)_2$	NO_2
27	$CONHCH_2$-phenyl	NO_2	77	CHC(CN)-phenylsulfonyl	NO_2
28	CHCHCO-(4-methoxyphenyl)	NO_2	78	$(CH_2)_4COOH$	Cl
29	CHCHCO-(4-nitrophenyl)	NO_2	79	CONH-(1-adamantyl)	Cl
30	CHCHCO-(2-carboxyphenyl)	NO_2	80	COOH	CN
31	CHCHCO-(4-chlorophenyl)	NO_2	81	$COO(CH_2)_3CH_3$	CN
32	CHCHCO-(3,4-dichlorophenyl)	NO_2	82	CONH-(1-adamantyl)	CN
33	CHCHCO-(3,4-dimethoxyphenyl)	NO_2	83	CHCHCO-phenyl	CF_3
34	CHCHCO-(4-hydroxy-3-methoxyphenyl)	NO_2	84	CHO	CN
35	CHCHCO-(4-hydroxyphenyl)	NO_2	85	CHO	CHO
36	$CHC(CN)COOCH_2CH_3$	NO_2	86	CHO	CF_3
37	$CHC(COCH_3)CH_2COOCH_3$	NO_2	87	$CHC(COCH_3)_2$	CF_3
38	$CONHCH_2$-(2-pyridyl)	NO_2	88	$CONHCH_2$-phenyl	CN
39	$CH_2CH(CN)CH_2OH$	NO_2	89	Cl	SO_2N-$(CH_2CH_3)_2$
40	$CH_2CH(CN)COOCH_2CH_3$	NO_2			
41	$CONHCH_2$-(2,4,6-trimethoxyphenyl)	NO_2	90	CN	CN
42	CH_2OH	NO_2	91	CHO	SO_2CH_3
43	$CH_2OCH_2CH_2OCH_3$	NO_2	92	Cl	SO_2CH_3
44	CH_2SCH_2COOH	NO_2	93	NO_2	F
45	$CHC(CN)CONH_2$	NO_2	94	$CH_2CH(CH_3)N(CH_3)CH_2CCH$	NO_2
46	CH_2-(2-pyrrolyl)	NO_2	95	CO-(4-methylphenyl)	NO_2
47	H	NO_2	96	$CH_2OCH_2CHCH_2O$	NO_2
48	Cl	NO_2	97	$CHC(CN)CON(CH_2CH_3)_2$	NO_2
49	$CH_2CH(COOH)NH_2$	NO_2	98	$COCH(CH_3)_2$	H
50	$CHC(CN)CON(CH_3)_2$	NO_2	99	$COOCH_2CH_2CH_3$	OH

Enzyme inhibition measurements

Biological activity was defined as the potency of a given compound to inhibit S-COMT isolated from rat brain. The activity was expressed as the negative logarithm of the 50% inhibitory concentration ($pIC_{50} = -\log(IC_{50})$). The determination of IC_{50} was performed by measuring the COMT activity in several drug concentrations. The experimental procedure follows the one used in the study on rat liver S-COMT inhibition [14].

Molecular descriptors

The quantum-chemical descriptors were calculated on a μVAX II computer with the AMPAC program (QCPE program no. 506) and the MOPAC program (QCPE program no. 455) using the semiempirical AM1 method [20]. Molecules were modelled by the CHEM-X program and all geometries were fully optimized [21]. Quantum-chemical descriptors were calculated for both neutral and anionic forms of the inhibitors (the hydroxyl group at position 4 is deprotonated; see Fig. 1). The calculated descriptors are associated with the electronic properties of the molecules (partial atomic charges and superdelocalisabilities of ring carbon C4 and catechol oxygens, HOMO (the highest occupied molecular orbital) and LUMO (the lowest unoccupied molecular orbital) energies, bond orders of C4-O bond etc.). To calculate these local electronic properties of compound 23, the following R_1 substituent was used: $R_1 = CH_2CH_2CH_2CH_2COOCH_3$, because the number of atoms in compound 23 is 73 and the maximum number of atoms allowed with MOPAC is 60. This is a reasonable solution because the long aliphatic chain at the end of the R_1 substituent does not influence the electronics of the ring system. Van der Waals volumes of R_1 and R_5 substituents were used to describe the sizes of these substituents (the numerical value of Volume/100 was used) [21]. Molar refractivities (MR) (a global molecular property) and π values of R_1 substituents were obtained using the CLOGP program (the numerical value of MR/10 was used) [22]. Variables used to characterize the COMT inhibitors are summarized in Table 2. The total number of chemical structural indices is 19.

Data analytical methods

The statistical calculations were carried out on a μVAX II computer with the SIMCA (version 4.2 1990) package (PLS and PCA methods) and the SAS statistical program package (SAS Institute Inc., USA) (linear regression analysis).

The principal component analysis (PCA) was performed on the chemical structural data matrix in order to analyze the similarities and differences in information content in the set of 19 variables and to study relationships between compounds with respect to these variables. By examination of the variable loadings, it is possible to evaluate the contribution of each variable to the corresponding principal component [23]. These loadings show how much the variable participates in each principal component dimension of the model. A large loading value (positive or negative) means that the variable is important in the model and a small loading value that it is only a minor contributor to the model [16]. The crossvalidation criterion was applied to ensure the statistical validity of PCA [24,25].

The PLS method was used to build a correlation model between the molecular descriptors and the inhibition activity data [26]. The number of significant factors (PLS dimensions) also in the PLS models is estimated by crossvalidation [24,25]. Thus a balance is maintained between high

TABLE 2
VARIABLES USED TO CHARACTERIZE THE COMT INHIBITORS

Variable no.	Property
1	MR (molar refractivity)
2	V_1 (van der Waals volume of R_1 substituent)
3	π (π-value of R_1 substituent)
4	E_{LUMO} (the energy of the lowest unoccupied molecular orbital for neutral molecule)
5	E_{HOMO} (the energy of the highest occupied molecular orbital for neutral molecule)
6	V_5 (van der Waals volume of R_5 substituent)
7	BOA (bond order of C4-O bond for anionic species)
8	C4A (partial atomic charge at C4 for anionic species)
9	O4A (partial atomic charge at O4 for anionic species)
10	O3A (partial atomic charge at O3 for anionic species)
11	SDHC4A (superdelocalisability on HOMO orbital at C4 for anionic species)
12	SDHO4A (superdelocalisability on HOMO orbital at O4 for anionic species)
13	E_{HOMOA} (the energy of the highest occupied molecular orbital for anionic molecule)
14	E_{LUMOA} (the energy of the lowest unoccupied molecular orbital for anionic molecule)
15	Q (an indicator parameter describing the ionization state of R_1 substituents of the molecules)
16	SDLC5 (superdelocalisability on LUMO orbital at C5 for neutral species)
17	SDLR5 (superdelocalisability on LUMO orbital at R5 substituent for neutral species)
18	C4 (partial atomic charge at C4 for neutral species)
19	O4 (partial atomic charge at O4 for neutral species)

descriptive power and lack of fit of the model to the data. Also for the PLS model, the extent and sign of correlation between the descriptors and activity is evaluated by the variable loadings.

The relevance of each variable in the PLS model can be judged by the modelling power (mpow) which expresses the explained standard deviation per variable. The modelling power of a variable is defined as follows:

$$\text{mpow}_i = 1 - \frac{s_i}{s_{oi}}$$

with s_i = residual standard deviation (RSD) of variable (i) after extracting A components (A = number of model dimensions); s_{oi} = initial standard deviation of variable (i). A mpow_i value close to one indicates a high relevance of the variable (i) in the PLS model [27,28].

The residuals e_{ik} (containing the part of the chemical descriptor data not explained by the PLS model) after the A significant model dimensions can be used to calculate measures of fit to the PLS model for each compound. Geometrically, the residual standard deviation of an individual compound (i) (RSD_i) is directly interpretable as the distance between the *i*th compound and the PLS model. This compound RSD can be compared with the pooled RSD of the compounds in the training set to see whether compound (i) is an outlier [17]. The RSD for the PLS model measures the typical distance between the PLS class model (class = training set) and a compound belonging to the class. The RSD ($=RSD_o$) for the PLS model and the RSD_i for the compound (i) are defined as follows [29]:

$$RSD_0 = \sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K} e_{ik}^2/(N-A-1)(K-A)} \quad ; \quad RSD_i = \sqrt{\sum_{k=1}^{K} e_{ik}^2/(K-A)}$$

with A = number of model dimensions in the PLS model; K = number of variables (k) (chemical descriptors); N = number of compounds (i).

To predict the activities of new analogues, their RSD values are compared to the RSD of the training set analogs, i.e. to the RSD for the PLS model. If the RSD of the new analogue is considerably larger than the RSD for the PLS model, this indicates that the structure of the new analogue differs significantly from the training set and the predicted activity values for this new compound are less reliable [26,30].

The variables were scaled to unit variance in order to give every variable the same influence in the data analysis [28].

RESULTS

Principal component analysis (PCA)

PCA was used in order to obtain an overview of the compound distribution in the molecular descriptor space. In addition, similarities and differences between variables can be detected. The descriptor matrix was modelled by five principal components (PC) which are significant according to crossvalidation and explain 76.4% of the descriptor data variance (PC1, 42.4%; PC2, 12.9%; PC3, 11.7%; PC4, 4.5%; PC5, 4.9%). The distribution of the compounds on the first three latent variables derived by PCA is shown in Fig. 2.

By examination of the loading plot, it is possible to evaluate the relationships between variables. Figure 3 displays the loading plot from the principal component analysis. Since the first three PCs explain a major part of the variance (67%), their loading vectors are plotted. According to the loading plot, variables C4 (no. 18) and C4A (no. 8) seem to contain similar information. Also variables BOA (no. 7), O4A (no. 9), O3A (no. 10), SDHC4A (no. 11), SDHO4A (no. 12) and SDLC5 (no. 16) are clustered in the loading plot. Similar behaviour is found in the variables MR (no. 1) and V_1 (no. 2). In general, the variables are widely distributed in the 3D space expanded by three principal components. Loading values (the correlation coefficients between variables and PCs) of the variables for the first three PCs are presented in Table 3.

The first PC is mostly influenced by electronic descriptors. It is notable that the size descriptors have a large contribution to the second component. In addition, the variable V_5 has a large and positive loading value in the third component, see Table 3. The variables with loading values greater than $|\pm 0.2|$ are most important in the model [16].

The PLS models

The relationship between inhibition activity and structural descriptors was analyzed by the PLS method [17]. The number of significant components was determined with crossvalidation by leaving one compound out of the analysis during each round [24, 25].

Norinder [30] has recently shown that an important factor in deriving a successful model (relationship) is the choice of initial compounds (the training set). In an experimental design-based

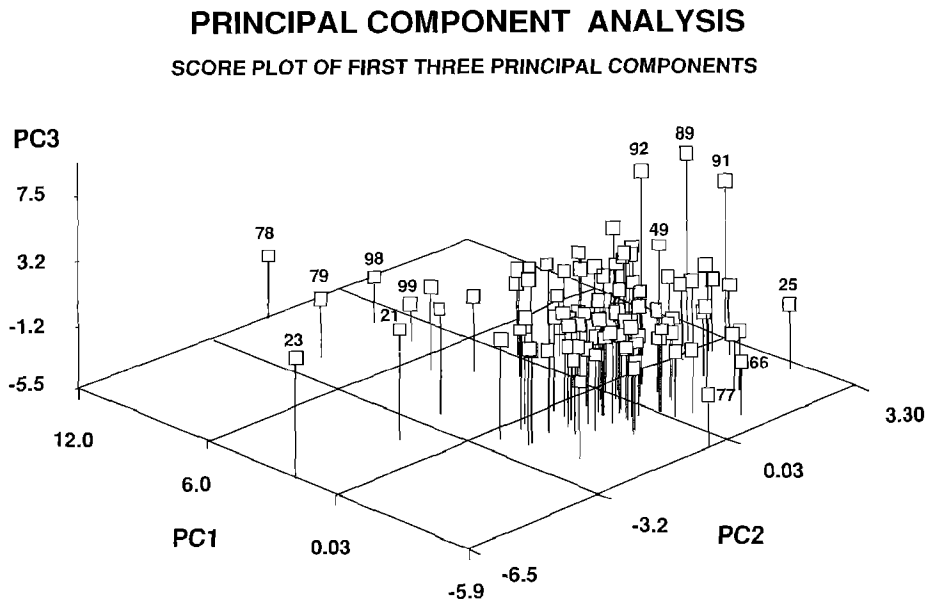

Fig. 2. Score plot of the first three principal components of the PCA of compounds (N = 99). PC1, PC2 and PC3 correspond to the scores of the principal components 1, 2 and 3, respectively (some borderline compounds are identified, Table 1).

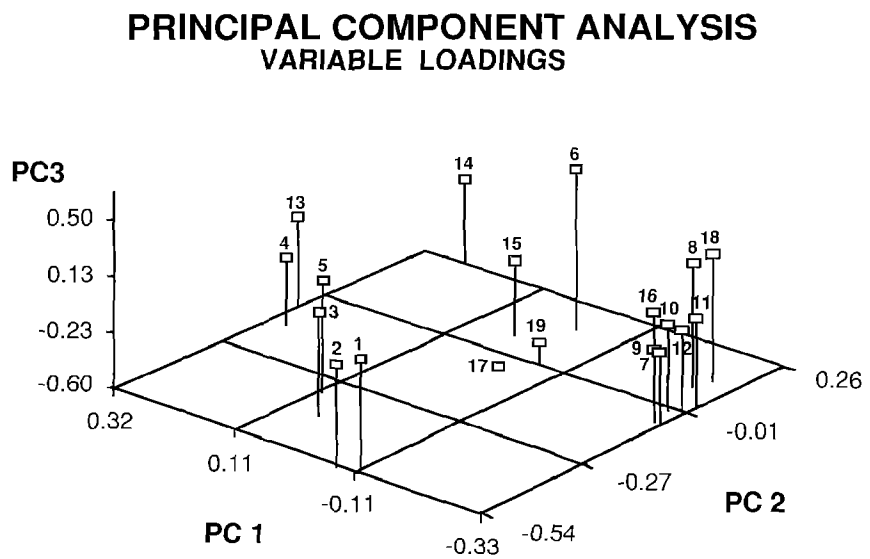

Fig. 3. Loading plot for the first three model dimensions of the PCA of compounds (N = 99) (the numbering of the variables is presented in Table 2).

TABLE 3
LOADINGS (p_{ak}) FROM PCA OF THE DESCRIPTOR MATRIX FOR COMT INHIBITORS (the first three PCs, a = 1, 2 and 3)

Variable (k)		p_{1k}	p_{2k}	p_{3k}	Variable (k)		p_{1k}	p_{2k}	p_{3k}
1	MR	−0.105	−0.523	0.103	11	SDHC4A	−0.326	0.021	−0.013
2	V_I	−0.078	−0.538	0.067	12	SDHO4A	−0.325	−0.014	−0.052
3	π	0.050	−0.410	0.080	13	E_{HOMOA}	0.325	−0.077	0.001
4	E_{LUMO}	0.294	−0.152	−0.150	14	E_{LUMOA}	0.251	0.256	−0.035
5	E_{HOMO}	0.099	−0.327	0.135	15	Q	0.031	0.069	−0.116
6	V_5	−0.024	0.155	0.460	16	SDLC5	−0.232	0.053	−0.132
7	BOA	−0.315	−0.060	−0.148	17	SDLR5	−0.026	−0.057	−0.573
8	C4A	−0.274	0.085	0.213	18	C4	−0.286	0.123	0.229
9	O4A	−0.310	−0.066	−0.125	19	O4	−0.058	0.001	−0.476
10	O3A	−0.302	−0.022	−0.021					

QSAR analysis the combined approach of a PCA-based selection of the training set compounds followed by a PLS analysis has been applied [30]. Following this type of approach the 32 compounds (model 1; see Table 1 and the Materials and Methods section) of the training set were selected on the basis of the PCA performed on the descriptor matrix. Geometrically, the compounds chosen as the training set were selected to span the descriptor space (hyperspace; 4D) as widely and efficiently as possible [30]. Since the first three PCs explain a major part of the data matrix variance, care was taken to efficiently cover the PC1-PC2-PC3 space.

The PLS analysis of the training set resulted in a significant (crossvalidated) two-component model (model 1) that described 69.1% (PLS1, 52.5%; PLS2, 16.6%) of the variance in the S-COMT inhibition activity data. The accumulated modelling power and PLS loadings are shown in Table 4. Variables with loading values greater than $|\pm 0.2|$ can be considered important in the model [16]. The first PLS component consists mainly of the variables of electronic nature. The second PLS component is influenced also by steric descriptors.

The residual standard deviation (RSD) for model 1 is 0.677. The RSDs of the individual compounds were less than or equal to two times that of the model except for the following compounds: 23, 78 and 93. Their RSD values were 2.12, 1.43 and 1.74, respectively. The calculated inhibition activity values of these analogues were more or less unreliable [26], particularly for compounds 23 and 93. Compound 23 has the largest R_1 substituent ($(CH_2)_4COO(CH_2)_{13}CH_3$), whereas compound 93 has a small R_5 substituent (F). The r^2 (explained) value for the plot of experimental vs. calculated inhibition activity values of the training set compounds is 0.715 and the r^2 value for the plot of experimental vs. calculated and predicted inhibition activity values of all (N = 99) compounds is 0.571, Fig. 4A. The observed (= experimental) and calculated (predicted) pIC_{50} values associated with the PLS models are summarized in Table 5.

For comparative purposes, a PLS analysis was also carried out on the full set of compounds. Inclusion of all 99 inhibitors into the PLS model (model 2) gave two significant components with an r^2 value for the plot of experimental vs. calculated inhibition activity values of 0.647. These PLS components described 64.7% (PLS1, 48%; PLS2, 16.7%) of the variance in the inhibition activity. The molecular descriptors that most influenced the PLS components are the same as in mo-

TABLE 4
PLS (MODEL 1) VARIABLE LOADINGS AND ACCUMULATED MODELLING POWER OF VARIABLES

Variables		Loadings		Accumulated mpow	
		load1	load2	mpow1	mpow2
1	MR	0.109	0.202	0.036	0.100
2	V_1	0.086	0.291	0.015	0.165
3	π	−0.061	0.168	0	0.033
4	E_{LUMO}	−0.256	0.325	0.313	0.673
5	E_{HOMO}	−0.100	0.174	0.026	0.067
6	V_5	−0.037	−0.437	0	0.394
7	BOA	0.338	0.046	0.773	0.787
8	C4A	0.238	−0.291	0.259	0.488
9	O4A	0.335	0.048	0.733	0.744
10	O3A	0.313	−0.148	0.561	0.653
11	SDHC4A	0.328	−0.106	0.666	0.726
12	SDHO4A	0.331	−0.128	0.697	0.813
13	E_{HOMOA}	−0.314	0.202	0.570	0.789
14	E_{LUMOA}	−0.255	0.090	0.323	0.333
15	Q	−0.003	−0.021	0	0
16	SDLC5	0.224	−0.111	0.224	0.240
17	SDLR5	0.081	0.375	0.012	0.292
18	C4	0.253	−0.311	0.303	0.609
19	O4	0.130	0.284	0.058	0.208

load1 = PLS loadings for the first PLS dimension; mpow1 = accumulated modelling power for the first PLS dimension.

del 1. Figures 5A and B display the loadings of the variables for the first PLS component of both models 1 and 2.

The results from model 2 are rather similar to the previous ones obtained by the PLS analysis of the training set (model 1). Compounds 23 and 93 still do not fit the PLS model. Their RSD values are (2.56 and 2.48, respectively) more than two times higher than that of model 2 (RSD = 0.694). This indicates that the reliability of calculated inhibition activity values of these analogues is not assured. In addition, compound 89 (RSD = 2.37) with the largest R_5 substituent ($SO_2N(C_2H_5)_2$) does not fit model 2.

Inspection of the results obtained from the PLS models 1 and 2 reveals that predicted (calculated) inhibition activity values for the compounds having a large R_1 substituent or having either a very small or large R_5 substituent are unreliable. In the case of model 1 a difference greater than ± 0.9 between experimental pIC_{50} and calculated (predicted) pIC_{50} was found for 14 compounds (Table 5). For model 2 the number of compounds was eight (Table 5). The predictive capability of the created PLS models 1 and 2 is not good. In addition, models 1 and 2 could not handle compound 49 with a zwitterionic R_1 substituent.

In order to improve the predictive power of the PLS models, two separate models were constructed by modifying the training set of model 1. The training set of model 3 consists of the compounds (N = 27, Table 1 and the Materials and Methods section) having as the R_5 substituent either NO_2, CN, CHO, Cl, F, OH, or H. The PLS analysis resulted in a significant (crossvalidated)

TABLE 5
THE EXPERIMENTAL pIC_{50} VALUES AND THE CALCULATED/PREDICTED pIC_{50} VALUES OBTAINED BY DIFFERENT PLS MODELS

Compound	Exp	M1	M2	M3	M4	D1	D2	D3	D4
1	7.07	6.11	6.45	6.73	6.34	0.96	0.62	0.34	0.73
2	7.43	7.05	7.13	6.99	6.81	0.38	0.30	0.44	0.62
3	8.22	7.44	7.70	7.69	7.76	0.78	0.52	0.53	0.46
4	7.57	6.73	7.26	7.62	7.66	0.84	0.31	−0.05	−0.09
5	7.48	6.49	6.98	7.37	7.39	0.99	0.50	0.11	0.09
6	7.64	6.48	7.00	7.34	7.24	1.16	0.64	0.30	0.40
7	7.92	6.81	7.29	7.49	7.29	1.11	0.63	0.43	0.63
8	7.70	7.61	7.72	7.53	7.58	0.09	−0.02	0.17	0.12
9	7.70	6.39	6.94	7.41	7.18	1.31	0.76	0.29	0.52
10	8.18	7.53	7.93	7.88	7.70	0.65	0.25	0.30	0.48
11	7.72	7.15	7.50	7.69	8.00	0.57	0.22	0.03	−0.28
12	7.40	7.53	7.67	7.54	7.65	−0.13	−0.27	−0.14	−0.25
13	7.00	6.76	7.03	7.29	7.48	0.24	−0.03	−0.29	−0.48
14	7.80	7.02	7.45	7.54	7.47	0.78	0.35	0.26	0.33
15	7.66	7.16	7.65	7.78	7.49	0.50	0.01	−0.12	0.17
16	7.80	6.87	7.17	7.35	7.37	0.93	0.63	0.45	0.43
17	7.62	6.61	7.03	7.26	7.09	1.01	0.59	0.36	0.53
18	7.68	7.43	7.68	7.67	7.67	0.25	0.00	0.01	0.01
19	7.92	7.46	7.86	7.94	7.65	0.46	0.06	−0.02	0.27
20	7.85	7.05	7.52	7.64	7.47	0.80	0.33	0.21	0.38
21	7.07	6.93	7.47	7.81	7.61	0.14	−0.40	−0.74	−0.54
22	7.52	6.83	7.10	7.27	7.46	0.69	0.42	0.25	0.06
23	6.00	7.16	8.01	8.37	7.44	−1.16	−2.01	−2.37	−1.44
24	7.82	7.19	7.63	7.74	7.43	0.63	0.19	0.08	0.39
25	7.92	7.29	7.48	7.51	7.82	0.63	0.44	0.41	0.10
26	7.70	7.07	7.56	7.86	7.73	0.63	0.14	−0.16	−0.03
27	7.74	7.16	7.49	7.68	7.58	0.58	0.25	0.06	0.16
28	7.89	7.22	7.67	7.74	7.54	0.67	0.22	0.15	0.35
29	7.92	7.83	7.94	7.77	7.56	0.09	−0.02	0.15	0.36
30	7.80	7.46	7.58	7.37	7.00	0.34	0.22	0.43	0.80
31	7.92	7.55	7.81	7.76	7.72	0.37	0.11	0.16	0.20
32	7.60	7.66	7.91	7.82	7.77	−0.06	−0.31	−0.22	−0.17
33	8.00	7.60	7.90	7.86	7.71	0.40	0.10	0.14	0.29
34	7.80	7.57	7.85	7.81	7.71	0.23	−0.05	−0.01	0.09
35	7.85	7.45	7.73	7.73	7.78	0.40	0.12	0.12	0.07
36	7.85	7.76	7.93	7.73	7.85	0.09	−0.08	0.12	0.00
37	7.80	6.97	7.43	7.59	7.43	0.83	0.37	0.21	0.37
38	7.28	7.13	7.43	7.61	7.68	0.15	−0.15	−0.33	−0.40
39	6.89	6.37	6.80	7.15	7.24	0.52	0.09	−0.26	−0.35
40	7.30	6.56	7.03	7.37	7.36	0.74	0.27	−0.07	−0.06
41	7.82	7.07	7.59	7.85	7.50	0.75	0.23	−0.03	0.32
42	6.76	6.20	6.65	7.00	7.12	0.56	0.11	−0.24	−0.36
43	7.07	6.24	6.74	7.15	7.10	0.83	0.33	−0.08	−0.03
44	6.92	6.49	6.78	6.84	6.81	0.43	0.14	0.08	0.11
45	7.74	7.22	7.59	7.58	7.42	0.52	0.15	0.16	0.32
46	7.15	6.42	6.96	7.27	7.38	0.73	0.19	−0.12	−0.23
47	6.85	5.89	6.37	6,87	6.95	0.96	0.48	−0.02	−0.10
48	7.60	6.25	6.71	7.07	7.35	1.35	0.89	0.53	0.25
49	5.49	6.22	6.60	7.13	7.33	−0.73	−1.11	−1.64	−1.84

TABLE 5 (continued)

Compound	Exp	M1	M2	M3	M4	D1	D2	D3	D4
50	7.74	7.54	7.76	7.66	7.84	0.20	−0.02	0.08	−0.10
51	7.85	7.64	7.88	7.77	7.79	0.21	−0.03	0.08	0.06
52	7.72	7.42	7.84	7.82	7.62	0.30	−0.12	−0.10	0.10
53	8.00	7.38	7.79	7.81	7.70	0.62	0.21	0.19	0.30
54	8.00	7.86	8.08	7.87	7.96	0.14	−0.08	0.13	0.04
55	8.00	7.47	7.82	7.77	7.58	0.53	0.18	0.23	0.42
56	7.10	6.95	7.21	7.37	7.44	0.15	−0.11	−0.27	−0.34
57	6.21	6.75	6.87	6.85	6.70	−0.54	−0.66	−0.64	−0.49
58	7.80	7.22	7.69	7.85	7.51	0.58	0.11	−0.05	0.29
59	7.54	6.89	7.36	7.54	7.26	0.65	0.18	0.00	0.28
60	7.72	7.45	7.77	7.60	7.29	0.27	−0.05	0.12	0.43
61	7.92	7.66	7.92	7.82	7.83	0.26	0.00	0.10	0.09
62	7.85	7.65	7.90	7.83	8.15	0.20	−0.05	0.02	−0.30
63	7.85	7.74	7.90	7.72	7.70	0.11	−0.05	0.13	0.15
64	7.85	7.73	7.98	7.85	7.75	0.12	−0.13	0.00	0.10
65	7.04	7.17	7.52	7.61	7.55	−0.13	−0.48	−0.57	−0.51
66	6.60	7.75	7.72	7.28	6.97	−1.15	−1.12	−0.68	−0.37
67	7.62	7.13	7.58	7.68	7.46	0.49	0.04	−0.06	0.16
68	7.80	7.30	7.55	7.60	7.74	0.50	0.25	0.20	0.06
69	7.46	7.12	7.54	7.68	7.52	0.34	−0.08	−0.22	−0.06
70	7.57	7.51	7.79	7.79	7.73	0.06	−0.22	−0.22	−0.16
71	7.92	7.50	7.77	7.69	7.67	0.42	0.15	0.23	0.25
72	7.80	7.23	7.54	7.64	7.58	0.57	0.26	0.16	0.22
73	7.33	7.60	7.93	7.94	7.89	−0.27	−0.60	−0.61	−0.56
74	7.51	7.48	7.95	8.00	7.86	0.03	−0.44	−0.49	−0.35
75	7.77	8.14	8.36	8.09	8.13	−0.37	−0.59	−0.32	−0.36
76	7.85	7.24	7.66	7.71	7.51	0.61	0.19	0.14	0.34
77	7.95	8.37	8.46	8.01	8.20	−0.42	−0.51	−0.06	−0.25
78	5.10	5.07	5.24	5.23	6.25	0.03	−0.14	−0.13	−1.15
79	6.40	6.52	6.83	6.53	8.13	−0.12	−0.43	−0.13	−1.73
80	5.70	6.49	6.47	6.51	6.15	−0.79	−0.77	−0.81	−0.45
81	7.72	6.50	6.76	6.97	7.25	1.22	0.96	0.75	0.47
82	7.04	6.58	6.94	7.17	7.46	0.46	0.10	−0.13	−0.42
83	6.59	6.95	6.97	7.33	6.35	−0.36	−0.38	−0.74	0.24
84	6.80	6.35	6.53	6.69	6.60	0.45	0.27	0.11	0.20
85	6.89	6.10	6.34	6.73	6.56	0.79	0.55	0.16	0.33
86	5.65	6.34	6.41	6.91	6.25	−0.69	−0.76	−1.26	−0.60
87	5.84	7.01	7.06	7.34	6.65	−1.17	−1.22	−1.50	−0.81
88	7.35	6.67	6.95	7.13	7.22	0.68	0.40	0.22	0.13
89	4.59	4.79	4.62	9.85	3.99	−0.20	−0.03	−5.26	0.60
90	7.46	6.28	6.47	6.64	6.88	1.18	0.99	0.82	0.58
91	4.70	5.20	5.65	8.31	4.96	−0.50	−0.95	−3.61	−0.26
92	4.80	4.82	5.27	8.03	5.24	−0.02	−0.47	−3.23	−0.44
93	6.48	6.91	6.87	5.93	8.91	−0.43	−0.39	0.55	−2.43
94	6.81	6.36	6.86	7.25	7.29	0.45	−0.05	−0.44	−0.48
95	7.61	6.91	7.41	7.62	7.36	0.70	0.20	−0.01	0.25
96	6.87	6.34	6.83	7.24	7.16	0.53	0.04	−0.37	−0.29
97	7.70	7.51	7.80	7.80	7.75	0.19	−0.10	−0.10	−0.05
98	5.22	5.81	6.18	5.52	7.26	−0.59	−0.96	−0.30	−2.04
99	5.70	6.33	6.48	5.94	6.88	−0.63	−0.78	−0.24	−1.18

Exp = experimental pIC_{50}; M1 = predicted/calculated pIC_{50} with model 1, D1 = Exp − M1, etc.

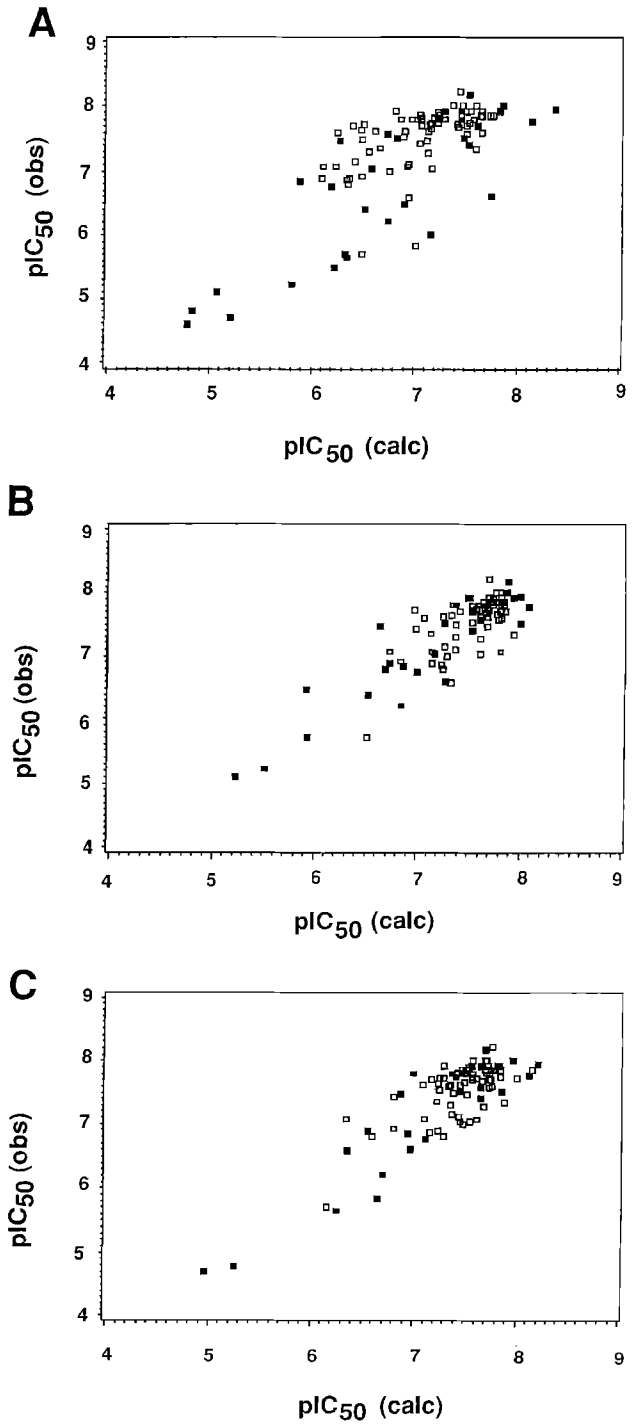

Fig. 4. Plot of observed (= experimental) and calculated/predicted pIC_{50} values for inhibition of rat brain S-COMT activity (filled squares are the compounds of the training set; open squares are predicted compounds): (A) model 1; (B) model 3; (C) model 4.

2-component model that described 82.2% (PLS1, 69.5%; PLS2, 12.7%) of the variance in the S-COMT inhibition activity data. The accumulated modelling power and PLS loadings are shown in Table 6. The first PLS component mainly consists of the electronic variables and the size descriptor V_5 and the second PLS component is mostly influenced by steric descriptors MR, V_1 and π and some electronic variables.

The residual standard deviation for model 3 is 0.650. According to their RSDs, the predicted pIC_{50} values of compounds 23, 89, 91 and 92 are more or less unreliable. Their RSD values were 2.87, 7.41, 3.14 and 3.21, respectively. Compounds 89, 91 and 92 have a large R_5 substituent (see Table 1). The r^2 value for the plot of experimental vs. calculated inhibition activity values of the training set compounds is 0.836 and the r^2 value for the plot of experimental vs. calculated and predicted inhibition activity values of the compounds (except compounds 23, 49, 83, 86, 87, 89, 91 and 92) is 0.729 (Fig. 4B). The predictive capability of model 3 is good. A difference greater than ± 0.8 between observed pIC_{50} and calculated (predicted) IC_{50} was found only for compounds 23, 49, 86, 87, 89, 91 and 92 which are already taken as outliers with respect to model 3.

The training set of model 4 consists of the compounds (N = 27, Table 1 and the Materials and Methods section) having either NO_2, CF_3, CHO, CN, SO_2CH_3, or $SO_2N(C_2H_5)_2$ as the R_5 substituent. The PLS analysis resulted in a significant (crossvalidated) 3-component model that described 83.3% (PLS1, 53.7%; PLS2, 19.2%; PLS3, 10.4%) of the variance in the inhibition activity data. The PLS loadings are shown in Table 7.

The residual standard deviation for model 4 is 0.651. The RSDs of compounds 23

TABLE 6
PLS (MODEL 3) VARIABLE LOADINGS AND ACCUMULATED MODELLING POWER OF VARIABLES

Variables		Loadings		Accumulated mpow	
		load1	load2	mpow1	mpow2
1	MR	0.125	0.390	0.063	0.276
2	V_1	0.086	0.419	0.018	0.256
3	π	−0.109	0.325	0.043	0.173
4	E_{LUMO}	−0.291	0.130	0.602	0.652
5	E_{HOMO}	−0.093	0.337	0.025	0.164
6	V_5	0.212	0.192	0.244	0.292
7	BOA	0.310	−0.082	0.801	0.846
8	C4A	0.280	0.070	0.525	0.528
9	O4A	0.306	−0.082	0.745	0.775
10	O3A	0.298	−0.047	0.660	0.661
11	SDHC4A	0.308	−0.022	0.767	0.765
12	SDHO4A	0.307	−0.071	0.762	0.785
13	E_{HOMOA}	−0.308	0.099	0.769	0.825
14	E_{LUMOA}	−0.249	0.069	0.370	0.367
15	Q	−0.010	−0.350	0	0.123
16	SDLC5	0.188	−0.327	0.181	0.348
17	SDLR5	−0.032	−0.286	0	0.072
18	C4	0.292	−0.003	0.611	0.603
19	O4	0.048	−0.218	0	0.032

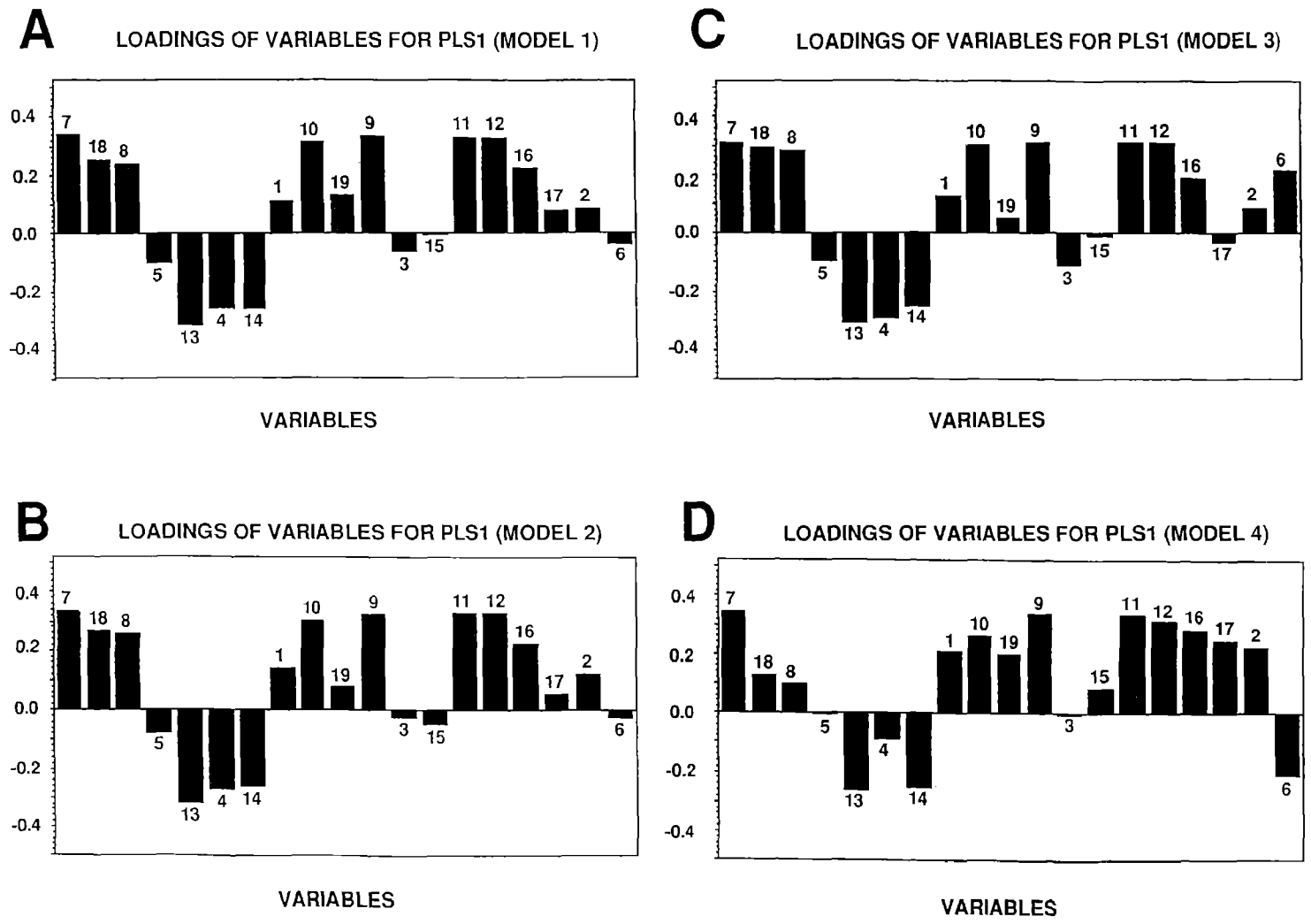

Fig. 5. Bar histograms of the variable loadings in the first PLS dimension of models: (A) model 1; (B) model 2; (C) model 3; (D) model 4.

TABLE 7
PLS (MODEL 4) VARIABLE LOADINGS

Variables		Loadings		
		load1	load2	load3
1	MR	0.210	0.025	−0.223
2	V_1	0.224	0.116	−0.181
3	π	−0.003	0.050	−0.061
4	E_{LUMO}	−0.090	0.399	−0.399
5	E_{HOMO}	−0.005	0.141	−0.307
6	V_5	−0.208	−0.310	−0.200
7	BOA	0.343	−0.036	0.133
8	C4A	0.098	−0.417	−0.179
9	O4A	0.338	−0.043	0.183
10	O3A	0.260	−0.274	0.196
11	SDHC4A	0.336	−0.120	0.130
12	SDHO4A	0.313	−0.204	−0.074
13	E_{HOMOA}	−0.261	0.303	−0.169
14	E_{LUMOA}	−0.251	0.226	0.226
15	Q	0.082	−0.135	−0.279
16	SDLC5	0.282	−0.212	−0.129
17	SDLR5	0.246	0.164	−0.325
18	C4	0.125	−0.385	0.373
19	O4	0.201	0.137	−0.237

(RSD = 3.67), 78 (RSD = 3.84), 79 (RSD = 3.05), 93 (RSD = 3.48), 98 (RSD = 2.75) and 99 (RSD = 2.01) are much higher than that of the model class. Compound 23 has the largest R_1 substituent whereas compounds 78, 79, 93, 98 and 99 have a small R_5 substituent (see Table 1). The r^2 value for the plot of experimental vs. calculated inhibition activity values of the training set compounds is 0.853 and the r^2 value for the compounds (except compounds 23, 49, 78, 79, 93, 98 and 99) is 0.762 (Fig. 4C). The predictive capability of model 4 is also good. A difference greater than ± 0.8 between experimental pIC_{50} and calculated (predicted) IC_{50} was found only for compounds 23, 49, 78, 79, 93, 98 and 99 which are already taken as outliers with respect to model 4 (Table 5).

Using the created PLS models 3 and 4, it is possible to obtain sound predicted inhibition activity values for all compounds except compounds 23 and 49. Compound 23 has the largest R_1 substituent ($(CH_2)_4COO(CH_2)_{13}CH_3$) and all the PLS models described above overestimate its pIC_{50} value. The R_1 substituent of compound 49 has a zwitterionic nature.

For comparative purposes two additional PLS models were constructed. First, all compounds having as the R_5 substituent either NO_2, CN, CHO, Cl, F, OH, or H were added to the training set (except compounds 23 and 49) (see model 3). The PLS analysis gave a significant (crossvalidated) 2-component model that described 74.7% (63.2 + 11.5%) of the variance in the inhibition activity data. The r^2 value for the plot of experimental vs. calculated inhibition activity values of the training set compounds is 0.739. A difference greater than $\pm$ 0.8 between experimental pIC_{50} and calculated pIC_{50} was found only for compound 65. In the second model the training set consisted of all compounds having either NO_2, CF_3, CHO, CN, SO_2CH_3, or $SO_2N(C_2H_5)_2$ as the R_5 substi-

tuent (except compounds 23 and 49) (see model 4). In this case a crossvalidated 2-component PLS model described 69.8% (46.8 + 23%) of the variance in the inhibition activity data with an r^2 value of 0.705. A difference greater than ± 0.8 between experimental pIC_{50} and calculated pIC_{50} was found for compounds 21, 66, 80, 86 and 87. Although the predictive capability of these PLS models is quite good, they explained a smaller part of the variance in the inhibition activity than models 3 and 4.

Because of the need of two separate PLS models (models 3 and 4), one explaining the pIC_{50} of the compounds having the smaller R_5 substituents and the other explaining the pIC_{50} of the compounds with the larger R_5 substituents compared to the size of an NO_2 substituent, it is obvious that S-COMT inhibition activity is nonlinearly related to the size of the R_5 substituent as is the case also with the R_1 substituent.

The above PLS modelling was based on linear terms of variables. Therefore, the descriptor matrix was expanded with the quadratic terms of the variables π, V_1 and V_5 in order to try to model nonlinear behaviour. Model 1 and also the other PLS models developed overestimate the pIC_{50} value of compound 23 (Table 5). The addition of V_1^2 or V_5^2, both V_1^2 and V_5^2, or π^2, V_1^2 and V_5^2 together, with the previously used descriptors (Table 2) to PLS modelling of the whole set of compounds still leads to models which overestimate the pIC_{50} of compound 23. The r^2 for the plots of experimental vs. calculated/predicted inhibition activity values of the compounds were not statistically better than in the case of model 2. On the other hand, neither could these models handle molecules having the smallest and largest R_5 substituents inside the same model.

Inclusion of the π^2 term to the PLS model of the whole set of compounds gave a significant (crossvalidated) 3-component PLS model that described 77.7% (PLS1, 48.8%; PLS2, 20%; PLS3, 8.9%) of the variance in the inhibition activity data. The r^2 value was 0.754. However, the model underestimates the pIC_{50} value (4.92) of compound 23. The RSD of the model was 0.667 and that of compound 23 was 2.79. The quadratic terms of π and V_1 resulted in a significant PLS model (four components) that described 80% (PLS1, 49%; PLS2, 19.2%; PLS3, 7.9%; PLS4, 3.9%) of the variance in the inhibition activity. In this case the r^2 value for the plot of experimental vs. calculated inhibition activity value of the compounds was 0.781. The calculated pIC_{50} value of compound 23 was 5.88 (difference between experimental and calculated was only 0.12). However, the RSD for compound 23 (2.46) was large compared with the RSD of the PLS model (0.697). This indicates that the predicted value may be unreliable. The accumulated modelling power of the variables π^2 and V_1^2 was 0.330 and 0.697, respectively. The inclusion of the quadratic terms of π and V_5 produced a significant PLS model (3-component model) that described 77.2% (PLS1, 52.2%; PLS2, 16.1%; PLS3, 8.9%) of the variance in the inhibition activity data which estimates the pIC_{50} (5.39) value of compound 23 close to its experimental value (6.00). Also in this case, the RSD for compound 23 (3.46) was much larger than the RSD of the model (0.681). The r^2 value for the plot of experimental vs. calculated inhibition activity value of the compounds was 0.748. The accumulated modelling power of the variables π^2 and V_5^2 was 0.150 and 0.473, respectively. Although the differences between the experimental and calculated pIC_{50} values are smaller for the latter two PLS models compared with model 2, there are compounds with a difference between the experimental and calculated pIC_{50} greater than ± 0.9 (for instance compounds 49, 81, 90 and 91). None of these models can estimate the pIC_{50} value for the zwitterionic compound 49.

The inclusion of the quadratic terms of π, V_1 or V_5 in the descriptor matrix does not improve the predictive capability of model 1. The pIC_{50} of compound 23 is still overestimated.

DISCUSSION

The selection of structural descriptors here partly follows the previously postulated binding model [14]. In the present case it was not possible for instance to find σ^- values for all R_1 substituents from the literature because of the large variation in R_1 substituents. On the other hand, estimation of these σ^- values was not reasonable. Instead, calculated molecular descriptors associated with the electronics of the catecholic ring were used, for instance the calculated bond order (BOA) of the C4-O bond for the anionic forms of the molecules. The magnitude of these variables is greatly affected by the electronic nature of the R_1 and R_5 substituents. The electron-withdrawing effect of the substituents at R_1 and R_5 on the catechol ring increases the ionization of 4-OH, and thereby favours the inhibitory effect [14]. We have also found previously that the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and the normalized frontier electron density at position 5 may be important properties in order to explain the structure–activity relationship [14]. The electronic properties were calculated using the AM1 semiempirical method as reported previously [14,20]. The AM1 Hamiltonian was chosen because of its ability to correctly reproduce the planar geometry of nitrobenzene and benzaldehyde and to reproduce hydrogen bonds [14]. In the present investigation, size descriptors of R_1 and R_5 substituents were also included.

Although the number of compounds was considerably higher than the number of variables and biological data was univariate, the use of the PLS method was necessary in order to get statistically relevant results, because of the history of the inhibitors and of significant intercorrelations between calculated descriptor variables. The correlation analysis (Pearson correlation coefficients by the SAS program) was carried out within the variables, and high intercorrelations were found between the electronic variables (e.g. BOA and O4A) and also between size descriptors (e.g. MR and V_1).

Principal component analysis was used for two purposes. Firstly to study relationships between compounds and descriptor variables, and secondly PCA was used to choose the compounds of the training set for PLS analysis. The descriptors with the same information content are clustered in the PC space, i.e. PCA can be used to detect intercorrelations among the molecular indices. For instance, Cocchi et al. [27] have used the PCA approach to choose the molecular variables for PLS analysis. They selected representative descriptors from each cluster [27]. In the present case, all the molecular descriptors were selected although some of the variables are clustered (Fig. 3). Irrelevant variables may introduce some noise in the PLS models. However, PLS models are less sensitive to this noise than the other regression methods [26].

Because the PLS models 1 and 2 have difficulties in predicting pIC_{50} values for the inhibitors having either a small or large R_5 substituent compared with the NO_2 substituent, two separate models were developed. Both new PLS models have a good predictive power, i.e. the predicted/calculated pIC_{50} match the experimental pIC_{50} (Figs. 4B and C). The PLS model 3 could not predict pIC_{50} values of the compounds having CF_3, SO_2CH_3, or $SO_2N(CH_2CH_3)_2$ as the R_5 substituent. On the other hand, the PLS model 4 could not incorporate the compounds having Cl, F, H, or OH as the R_5 substituent. The PLS models 3 and 4 reveal that S-COMT inhibition activity is nonlinearly related to the size of the R_5 substituent. This is also shown by the signs of the loading values of the variable V_5 in models 3 and 4 (Tables 6 and 7, Figs. 5C and D). In both derived models 3 and 4, the variable V_5 is important to the model. The PLS models 3 and 4 could not cor-

rectly predict the pIC_{50} values of compound 23 with the largest R_1 substituent and compound 49 with a zwitterionic R_1 substituent.

The following discussion of the relevance of the variables in the PLS modelling is based on model 3. It has been shown [14] that the electronic character of the R_1 and R_5 substituents is important for inhibition activity of inhibitors. However, in the case of model 4 all the R_5 substituents of the training set compounds have only electron-withdrawing nature, while in model 3 the R_5 substituents vary from strong electron-releasing substituents such as OH to strong electron-withdrawing substituents like NO_2 within the training set compounds. So it is obvious that the influence of the electronic properties of substituents towards the catecholic ring system observed by calculated descriptors and the importance of these calculated descriptors in the PLS models is more reliably and accurately included in model 3, although the predictive power of both PLS models 3 and 4 is good.

According to Cocchi et al. [27] theoretical indices that contribute to the PLS model are indicated by their loading and modelling power values. Table 6 summarizes the loading values and accumulated modelling power values of the variables in the PLS model 3. Large (>0.2) and positive loading values were found for the variable BOA and the descriptors with the same information content (calculated to anionic forms of inhibitors) in the first PLS component (Table 6). Also, large modelling power values indicate a high relevance of these variables in model 3 (Table 6). It was also found that the double bond nature of the C4-O bond was increased (the calculated value of BOA $\rightarrow$ 2) with increasing inhibition activity among inhibitors, which means that the electronic effects of R_1 and R_5 substituents towards the ring system stabilize the anionic forms of S-COMT inhibitors. The substituents that cause this effect have an electron-withdrawing nature. The importance of electronic properties of the R_1 and R_5 substituents is in accordance with the earlier binding model [14]. Also, the E_{LUMO} variable is relevant in model 3. It has a negative large loading value in the PLS1 dimension. It can be concluded that more negative LUMO orbital energy will lead to higher inhibition activity. The calculated HOMO and LUMO molecular orbital energies of anionic forms of inhibitors have large negative loading values as well. Also, these variables associated with the molecular orbital energies have large modelling power values, which reflect their importance in PLS models (Table 6).

In PLS model 3 the steric parameter V_5 has a large positive loading value, indicating that the inhibition activity is positively correlated to the size of the R_5 substituent up to the size of the NO_2 group. The size and lipophilic character of the R_1 substituent also seem to be important descriptors in model 3. The molecular descriptors V_1, π and also MR have large and positive loading values in the second PLS dimension. It can be concluded that the inhibition activity increases with increasing size of the R_1 substituent and its hydrophobic nature. However, there is a limit to the size of the R_1 substituent. Compound 23, having the largest R_1 substituent, is a nonpotent inhibitor. It is obvious that the R_1 side chain of compound 23 is too large to be fitted in the receptor active site. The modelling power values of these variables associated with the size and lipophilic character of inhibitors are considerably smaller compared with the mpow values for variables with electronic nature (Table 6). This shows the major contribution of variables describing electronic properties of inhibitors in PLS models.

It is notable that the variable Q has a large negative loading value in the same PLS dimension where the variable π is positively correlated to the PLS dimension. Q is an indicator parameter describing the ionization state of the R_1 substituent (the value of Q is 0 for neutral R_1 substituents).

With the CLOGP program it is possible to calculate log P values only for neutral molecules ($\pi = \log P_{RX} - \log P_{RH}$), so log P values of ionized compounds are calculated as neutral ones. From the loading value of Q it can be concluded that ionized R_1 substituents decrease the inhibition activity. The variables SDLC5 and SDLR5 contribute to model 3 with large negative loadings in the second PLS dimension. Earlier we have used normalized frontier electron density at position 5 (C5+R5) as a QSAR parameter (F_5^N) and found similar negative correlation to biological activity [14]. It was concluded [14] that the R_5 substituent could also be directly involved in an attractive interaction with the binding site, in addition, via its steric repulsion. The parameter used, F_5^N, is very similar to SDLC5 and SDLR5.

In the present work, linear two-block PLS models have been applied to model the relation between biological activity data and molecular descriptors [31]. Recently, so-called nonlinear PLS modelling [31,32] has been introduced to relate data matrices with a curved relationship. For instance, the analytical calibration curve may depart from linearity, and structure–activity relationships and the response surface used for optimization may have a maximum or a minimum in the domain investigated [31]. Wold et al. [31] have pointed out that experience and understanding of the properties of nonlinear models must be increased to make nonlinear PLS a routine method [31]. However, application of nonlinear models to our data would be an interesting continuation to our QSAR studies.

CONCLUSIONS

The structure–activity relationships between 99 soluble COMT inhibitors are modelled with two separate PLS models. The structure of inhibitor molecules is characterized by theoretical molecular descriptors. Model 3 could incorporate the compounds having smaller R_5 substituents compared to the NO_2 substituent and correspondingly model 4 could soundly predict pIC_{50} values for the compounds with larger R_5 substituents compared to the NO_2 group. The inspection of the variable loadings (Figs. 5C and D) derived from PLS models reveals that inhibition activity is nonlinearly related to the size of the R_5 substituent. The predictive capability of these PLS models is good.

A number of conclusions can be drawn from the structure–activity relationships. First, the electronic properties of the R_1 and R_5 substituents are very important with respect to increased binding affinity. An inspection of the modelling power values in Table 6 shows the priority of the electronic effect and lower contribution of for instance size descriptors. The substituents that increase inhibition activity have an electron-withdrawing nature. Correspondingly, it was found earlier that electron-withdrawing substituents at positions 1 and 5 increased the inhibition activity [14]. Secondly, the PLS analysis shows that the inhibition activity is dependent on the size of the R_5 substituent. The inhibition activity increased with increased van der Waals volume of the R_5 substituent up to approximately the size of the NO_2 group. So it can be concluded that the size of the NO_2 group as an R_5 substituent is near optimal. Thirdly, inhibition activity is positively correlated to the size and lipophilicity of the R_1 substituent. However, there is a limit to the size of the R_1 substituent. Compound 23, having the largest R_1 substituent, is a nonpotent inhibitor. The PLS models could not correctly predict the pIC_{50} value for compound 23. Finally, ionized R_1 substituents decrease the inhibition activity. The pIC_{50} value of compound 49 having an R_1 substituent with a zwitterionic nature could not be predicted correctly with derived PLS models. So only two

compounds, 23 and 49 of the whole set of compounds (N = 99) can be taken as outliers in PLS modelling. The derived PLS models are well in accordance with the previously obtained linear regression equations [14].

Supplementary material available. Descriptor data can be obtained from the authors on request.

REFERENCES

1 Axelrod, J. and Tomchick, R., J. Biol. Chem., 233 (1958) 702.
2 Guldberg, H. and Marsden, C., Pharmacol. Rev., 27 (1975) 135.
3 Ball, P., Knuppen, R., Haupt, M. and Breuer, H., J. Clin. Endocrinol., 34 (1972) 736.
4 Borchardt, R.T., In Jakoby, W.B. (Ed.), Enzymatic Basis of Detoxification, Vol. II, Academic Press, New York, 1980, p. 43.
5 Assicot, M. and Bohuon, C., Biochimic, 52 (1971) 871.
6 Nissinen, E., Biochem. Pharmacol., 33 (1984) 3105.
7 Salminen, M., Lundström, K., Tilgmann, C., Savolainen, R., Kalkkinen, N. and Ulmanen, I., Gene, 93 (1990) 241.
8 Tilgmann, C. and Kalkkinen, N., FEBS Lett., 264 (1990) 95.
9 Bertocci, B., Miggiano, V., Da Prada, M., Dembic, Z., Lahm, H.-W. and Malherbe, P., Proc. Natl. Acad. Sci. USA, 88 (1991) 1416.
10 Lundström, K., Salminen, M., Jalanko, A., Savolainen, R. and Ulmanen, I., DNA Cell Biol., 10 (1991) 181.
11 Linden, I.-B., Nissinen, E., Etemadzadeh, E., Kaakkola, S., Männistö, P. and Pohto, P., J. Pharmacol. Exp. Ther., 247 (1988) 289.
12 Männistö, P. and Kaakkola, S., Trends Pharmacol. Sci., 10 (1989) 54.
13 Bäckström, R., Honkanen, E., Pippuri, A., Kairisalo, P., Pystynen, J., Heinola, K., Nissinen, E., Linden, I.-B., Männistö, P., Kaakkola, S. and Pohto, P., J. Med. Chem., 32 (1989) 841.
14 Taskinen, J., Vidgren, J., Ovaska, M., Bäckström, R., Pippuri, A. and Nissinen, E., Quant. Struct.-Act. Relat., 8 (1989) 210.
15 Dunn III, W.J., Wold, S., Edlund, U., Hellberg, S. and Gasteiger, J., Quant. Struct.-Act. Relat., 3 (1984) 131.
16 Hellberg, S., Wold, S., Dunn III, W.J., Gasteiger, J. and Hutchings, M.G., Quant. Struct.-Act. Relat., 4 (1985) 1.
17 Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. and Sjöström, M., In Kowalski, B.R. (Ed.), Chemometrics-Mathematics and Statistics in Chemistry, NATO ASI Series C No. 138, Reidel Publ. Co., Dordrecht, 1984, pp. 17–95.
18 Cruciani, G., Baroni, M., Bonelli, D., Clementi, S., Ebert, C. and Skagerberg, B., Quant. Struct.-Act. Relat., 9 (1990) 101.
19 Jolliffe, I.T., Principal Component Analysis, Springer, New York, 1986.
20 Dewar, M.J.S., Zoebisch, E.G., Eamonn, F.H. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
21 CHEM-X, developed and distributed by Chemical Design Ltd., Oxford, UK.
22 Daylight Chemical Information Systems Inc., MedChem Software, Release 3.54, 1990.
23 Lindgren, F., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M. and Wold, S., Quant. Struct.-Act. Relat., 10 (1991) 36.
24 Wold, S., Technometrics, 20 (1978) 397.
25 Cramer III, R.D., Bunce, J.D., Patterson, D.E. and Frank, I.E., Quant. Struct.-Act. Relat., 7 (1988) 18.
26 Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S., J. Med. Chem., 30 (1987) 1126.
27 Cocchi, M., Menziani, M.C., Rastelli, G. and De Benedetti, P.G., Quant. Struct.-Act. Relat., 9 (1990) 340.
28 SIMCA-R Manual, Multivariate Data Analysis version 4.2, 1990 (UMETRI AB, Sweden).
29 Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L., Chemometrics: A Textbook. Data Handling in Science and Technology, Vol. 2, Elsevier, Amsterdam, 1988, pp. 403-407.
30 Norinder, U., J. Comput.-Aided Mol. Design., 4 (1990) 381.
31 Wold, S., Kettaneh-Wold, N. and Skagerberg, B., Chemometrics Intel. Lab. Sys., 7 (1989) 53.
32 Frank, I.E., Chemometrics Intel. Lab. Sys., 8 (1990) 109.