Journal of Computer-Aided Molecular Design, 3 (1989) 111–132
ESCOM

J-CAMD 066

Pattern recognition study of QSAR substituent descriptors

Han van de Waterbeemd*, Nabil El Tayar, Pierre-Alain Carrupt and Bernard Testa**

School of Pharmacy, University of Lausanne, Place du Château 3, CH-1005 Lausanne, Switzerland

Received 12 October 1988
Accepted 4 April 1989

Key words: QSAR parameters; Principal component analysis; Cluster analysis; Drug design

SUMMARY

Parameter values for 59 common substituents and 74 descriptors used in QSAR studies were compiled. This data matrix was analysed by a variety of multivariate techniques. Linear regression confirmed that lipophilicity can be factorized into two terms, one related to molecular bulk and the other to polarity. Principal component analysis (PCA) of parameters revealed 5 significant principal components and a grouping of lipophilic, steric and electronic parameters. The different loadings of parameters with 5 PCA were also explored. The classification of substituents by cluster analysis (CA) proved rather disappointing. In contrast, the SIMCA method classified substituents of increasing bulk into 5 groups of increasing polarity.

INTRODUCTION

Since the introduction of the concept of quantitative structure–activity relationship (QSAR) studies in drug design by Hansch and coworkers in the early 1960s [1–4], medicinal chemists have made many proposals to describe molecular structure in a quantitative way. Some of us have recently compiled a great number of important and less important molecular (global) and substituent (local) descriptors or parameters which have been quoted in the literature over the last 25 years [5]. Traditionally QSAR parameters are classified in 3 groups, namely lipophilic, steric and electronic ones. In our compilation [5] we have in addition proposed solubility-related and formula-derived parameters. To this latter class belong, for example, molecular connectivity and related indices. However, various descriptors remain which cannot readily be classified, e.g. indicator or dummy variables. The great diversity of potentially useful descriptors is quite puzzling and one may wonder about the redundancy of the information encoded in the various descriptors.

*Present address: Hoffmann-La Roche AG, CH-4002 Basel, Switzerland.
**To whom correspondence should be addressed.

Statistical methods have been used to unravel the overlap or intercorrelation between parameters. Indeed this should be a crucial step in initiating multiple linear regression (MLR) studies [5, 6]. Correlations between lipophilic and steric parameters [7], or between molecular connectivity and steric parameters [8], have been studied extensively by linear regression. Pattern recognition techniques such as factor analysis (FA) [7], principal component analysis (PCA) [9], and cluster analysis (CA) [10], are more global and informative approaches. Thus an analysis of 18 QSAR parameters revealed 4 different clusters representing δ-inductive or field effects, π-resonance effects, steric terms and a group of ill-defined effects [10]. The study of Alunni et al. [9] involved a set of 7 descriptors and 28 substituents and confirmed that certain descriptors are strongly correlated, i.e. they have comparable information content. The strong grouping among the substituents into 4 separate classes, namely alkyls, H-bond donors, H-bond acceptors and halogens, is also interesting. Clustering behaviour has also been studied by the Pomona College group [2, 3]. Such

TABLE I
QSAR SUBSTITUENT DESCRIPTORS

No.	Code	Symbol	Description
1	PIAR	π_{ar}	aromatic substituent constant [1, 2, 4, 12, 13]
2	PIAL	π_{al}	aliphatic substituent constant [2, 12]
3	FARR	f_{ar}	aromatic hydrophobic fragmental constant [14, 15]
4	FALR	f_{al}	aliphatic hydrophobic fragmental constant [14, 15]
5	FARHL	f_{ar}'	aromatic lipophilic fragmental constant [2, 12]
6	FALHL	f_{al}'	aliphatic lipophilic fragmental constant [2, 12]
7	K	κ	electron donor-acceptor constant [16]
8	ES	E_S	Taft steric parameter [2, 13, 17, 18]
9	ESC	E_S^C	corrected Taft steric parameter [2, 19, 20]
10	UPS	υ	upsilon steric parameter [17, 21, 22]
11	VW	V_W	van der Waals volume [23–26]
12	VTSAR	V_{ar}^{TS}	fragmental volume aromatic substituent [27]
13	VTSAL	V_{al}^{TS}	fragmental volume aliphatic substituent [27]
14	MR	MR	molar refractivity [2, 12, 13, 19, 28, 29]
15	PR	Pr	parachor [30–33]
16	L	L	2nd generation STERIMOL length parameter [34, 35]
17	B1	B_1	2nd generation STERIMOL minimum width [34, 35]
18	B5	B_5	2nd generation STERIMOL maximum width [34, 35]
19	SMHL	σ_m^{HL}	Hammett constant for meta substitution [2, 12]
20	SPHL	σ_p^{HL}	Hammett constant for para substitution [2, 12]
21	SMSW	σ_m^{SW}	Hammett constant for meta substition [36]
22	SPSW	σ_p^{SW}	Hammett constant for para substitution [36]
23	SMM	σ_m^{M}	Hammett constant for meta substitution [37]
24	SPM	σ_p^{M}	Hammett constant for para substitution [37]
25	SME	σ_m^{E}	Hammett constant for meta substitution [17]
26	SPE	σ_p^{E}	Hammett constant for para substitution [17]
27	SSTAR	σ^*	electronic effect parameter for aliphatic substituents [13, 38, 39]
28	SIM	σ_I^{M}	Hammett constant for inductive effects [37]
29	SRM	σ_R^{M}	Hammett constant for resonance effects [37]
30	SICH	σ_I^{CH}	electronic parameter for inductive effects [40]
31	SRCH	σ_R^{CH}	electronic parameter for resonance effects [40]

TABLE I continued
QSAR SUBSTITUENT DESCRIPTORS

No.	Code	Symbol	Description
32	SIBR	σ_I^{BR}	electronic parameter for inductive effects [41]
33	F	F	field parameter [2, 42]
34	R	R	resonance parameter [2, 42]
35	FNEW	F'	corrected field parameter [43]
36	RNEW	R'	corrected resonance parameter [43]
37	SF	σ_F	inductive field parameter [44]
38	SX	σ_X	substituent electronegativity constant [45]
39	IOTA	ι	inductive substituent parameter [46, 47]
40	SSAR	σ_S^{ar}	entropy constant for aromatic group [48]
41	SSAL	σ_S^{al}	entropy constant for aliphatic group [48]
42	E	E	electronic substituent parameter based on SOMO-SOMO energy difference [49]
43	RE	RE	electronic parameter from path resistivity [49]
44	I	I	electronic parameter from flow intensity [49]
45	MUAR	μ_{ar}	aromatic group dipole moment [26, 50]
46	MUAL	μ_{al}	aliphatic group dipole moment [39]
47	LAMDAR	Λ_{ar}	lipophobic constant for aromatic substituent [27]
48	LAMDAL	Λ_{al}	lipophobic constant for aliphatic substituent [27]
49	HA	HA	hydrogen-bond acceptor indicator [2]
50	HD	HD	hydrogen-bond donor indicator [2]
51	HB	HB	hydrogen-bond parameter representing number of atoms capable of forming a H-bond [26]
52	X0AR	${}^0\chi_{ar}$	0th order molecular connectivity for aromatic group [51]
53	X0VAR	${}^0\chi_{ar}^v$	0th order valence molecular connectivity for aromatic group [51]
54	X1AR	${}^1\chi_{ar}$	1st order molecular connectivity for aromatic group [51]
55	X1VAR	${}^1\chi_{ar}^v$	1st order valence molecular connectivity for aromatic group [51]
56	SB	S_b	steric branching [52]
57	B	B	bulk parameter from BC(DEF) set [53]
58	C	C	cohesiveness parameter from BC(DEF) set [53]
59	RAND	RAND	random numbers between 0 and 1
60	MW	MW	molecular weight
61	B1OLD	B_1	1st generation STERIMOL minimum width [54]
62	B2	B_2	STERIMOL width parameter [54]
63	B3	B_3	STERIMOL width parameter [54]
64	B4	B_4	STERIMOL width parameter [54]
65	LOLD	L	1st generation STERIMOL length parameter [54]
66	NA	NA	number of atoms in substituent
67	NHET	NHET	number of hetero atoms in substituent
68	NC	NC	number of carbons in substituent
69	NHYDR	NHYDR	number of hydrogen atoms in substituent
70	ESK	E_S^K	E_S calculated from Kier's kappa values [55, 56]
71	K0	${}^0\kappa$	zero-order molecular shape parameter, coding for the uniqueness of the atoms [55, 56]
72	K1	${}^1\kappa$	first-order molecular shape parameter, related to the complexity of the molecule [55, 56]
73	K3	${}^3\kappa$	third-order molecular shape parameter, encoding for the centrality of branching [55, 56]
74	KSI	Ξ	Kier steric index [55]

TABLE 2
PARAMETER VALUES OF DESCRIPTORS 59–74 FOR 59 COMMON SUBSTITUENTS

STRUCT	RAND	MW	B1OLD	B2	B3	B4	LOLD	NA
Br	0.71	79.91	1.95	1.95	1.95	1.95	3.83	1
Cl	0.08	35.45	1.80	1.80	1.80	1.80	3.52	1
F	0.73	19.00	1.35	1.35	1.35	1.35	2.65	1
I	0.19	126.90	2.15	2.15	2.15	2.15	2.15	1
NO_2	0.01	46.01	1.70	1.70	2.44	2.44	3.44	3
H	0.84	1.01	1.00	1.00	1.00	1.00	2.06	1
OH	0.69	17.01	1.35	1.35	1.35	1.93	2.74	2
SH	0.70	33.07	1.70	1.70	1.70	2.33	3.47	2
NH_2	0.20	16.02	1.50	1.50	1.84	1.84	2.93	3
SO_2NH_2	0.24	80.09	2.11	2.67	2.67	3.07	3.82	6
CF_3	0.46	69.01	1.98	2.44	2.44	2.61	3.30	4
OCF_3	0.51	85.01	1.35	2.44	2.44	3.33	4.57	5
SO_2CF_3	0.46	133.07	2.11	2.67	2.67	3.64	4.11	7
SCF_3	0.05	101.07	1.70	2.44	2.44	3.69	4.89	5
CN	0.57	26.02	1.60	1.60	1.60	1.60	4.23	2
SCN	0.53	58.08	1.70	1.70	1.70	4.45	4.08	3
NCS	0.52	58.08	1.50	1.64	1.76	4.24	4.29	3
CHO	0.53	29.02	1.60	1.60	2.00	2.36	3.53	3
COOH	0.68	45.02	1.60	1.60	2.36	2.66	3.91	4
$CONH_2$	0.76	44.03	1.60	1.60	2.42	3.07	4.06	5
$OCONH_2$	0.11	60.03	1.35	1.60	1.60	3.62	4.82	6
CH_3	0.71	15.04	1.52	1.90	1.90	2.04	3.00	4
OCH_3	0.92	31.03	1.35	1.90	1.90	2.87	3.98	5
CH_2OH	0.70	31.03	1.52	1.90	1.90	2.70	3.97	5
$NHCONH_2$	0.88	59.05	1.84	1.84	1.94	3.61	5.09	7
SO_2CH_3	0.23	79.10	2.11	2.67	2.67	3.15	4.37	7
SCH_3	0.06	35.09	1.70	1.90	1.90	3.26	4.30	5
$NHCH_3$	0.60	30.05	1.50	1.90	1.90	3.08	3.53	6
C_2H	0.38	25.03	1.60	1.60	1.60	1.60	4.66	3
CH_2CN	0.85	27.05	1.52	1.90	1.90	4.12	3.99	5
C_2H_3	0.70	27.05	1.60	1.60	2.00	3.09	4.29	5
$COCH_3$	0.24	43.05	1.90	1.90	2.36	2.93	4.06	6
$COOCH_3$	0.60	59.05	1.90	1.90	2.36	3.36	4.85	7
$OCOCH_3$	0.04	59.05	1.35	1.90	1.90	3.68	4.87	7
CH_2COOH	0.24	59.05	1.60	2.16	2.35	3.75	4.69	7
OCH_2COOH	0.27	75.04	1.43	2.07	2.24	3.77	5.05	8
$NHCOCH_3$	0.33	58.06	1.50	1.90	1.94	3.61	5.15	8
$NHCOOCH_3$	0.26	74.06	1.53	2.37	2.62	4.10	5.79	9
C_2H_5	0.23	29.06	1.52	1.90	1.90	2.97	4.11	7
OC_2H_5	0.53	45.06	1.35	1.90	1.90	3.36	4.92	8
$N(CH_3)_2$	0.60	44.08	1.50	2.56	2.80	2.80	3.53	9
C_3H_5	0.86	41.07	1.98	2.24	2.29	2.88	4.14	8
$COOC_2H_5$	0.10	73.07	1.90	1.90	2.36	4.29	5.96	10
C_3H_7	0.61	43.09	1.52	1.90	1.90	3.49	5.05	10
$CH(CH_3)_2$	0.66	43.09	2.04	2.76	3.16	3.16	4.11	10
OC_3H_7	0.56	59.09	1.35	1.90	1.90	4.30	6.05	11
$OCH(CH_3)_2$	0.94	59.09	1.35	1.90	3.16	3.61	4.59	11
C_4H_9	0.20	57.12	1.52	1.90	1.90	4.42	6.17	13
$C(CH_3)_3$	0.20	57.12	2.59	2.86	2.86	2.97	4.11	13
OC_4H_9	0.09	73.12	1.35	1.90	1.90	4.79	6.99	14
NHC_4H_9	0.35	72.13	1.50	1.90	1.90	4.97	7.01	15
$N(C_2H_5)_2$	0.51	72.13	1.43	2.51	2.80	4.60	4.79	15
C_5H_{11}	0.67	71.14	1.52	1.90	1.90	4.94	7.11	16
C_6H_5	0.74	77.11	1.70	1.70	3.11	3.11	6.28	11
OC_6H_5	0.11	93.11	1.35	3.11	3.11	5.89	4.51	12
NHC_6H_5	0.11	92.12	1.50	3.11	3.11	5.95	4.53	13
C_6H_{11}	0.53	85.17	2.04	3.16	3.16	3.49	6.17	17
COC_6H_5	0.51	105.12	2.36	3.11	3.11	5.98	4.57	13
$CH_2CH_2C_6H_5$	0.43	105.16	1.52	3.11	3.11	3.16	8.33	17

TABLE 2 continued

PARAMETER VALUES OF DESCRIPTORS 59–74 FOR 59 COMMON SUBSTITUENTS

STRUCT	NHET	NC	NHYDR	ESK	K0	K1	K3	KSI
Br	1	0	0	0.14	0.000	1.480	1.200	1.76
Cl	1	0	0	0.01	0.000	1.290	1.010	1.57
F	1	0	0	–0.11	0.000	0.930	0.650	1.21
I	1	0	0	0.25	0.000	1.730	1.450	2.01
NO_2	3	0	0	0.84	0.829	2.880	1.380	3.55
H	0	0	1	–0.61	0.000	0.000	0.000	0.00
OH	1	0	1	–0.01	0.000	0.960	0.680	1.24
SH	1	0	1	0.01	0.000	1.350	1.070	1.63
NH_2	1	0	2	–0.01	0.000	0.960	0.680	1.24
SO_2NH_2	4	0	2	1.87	0.977	4.230	1.273	6.21
CF_3	3	1	0	1.63	0.977	3.790	0.969	5.63
OCF_3	4	1	0	0.80	2.063	4.750	4.327	3.11
SO_2CF_3	6	1	0	2.97	3.880	7.060	1.110	9.13
SCF_3	4	1	0	0.97	2.063	5.140	4.709	3.51
CN	1	1	0	–0.01	0.602	1.490	1.179	1.20
SCN	2	1	0	0.17	1.431	2.840	2.527	1.72
NCS	2	1	0	0.15	1.431	2.800	2.493	1.68
CHO	1	1	1	0.00	0.602	1.670	1.360	1.38
COOH	2	1	1	0.73	0.829	2.630	1.130	3.30
$CONH_2$	2	1	2	0.73	0.829	2.630	1.130	3.30
$OCONH_2$	3	1	2	0.41	1.806	3.670	3.285	2.25
CH_3	0	1	3	–0.01	0.000	0.000	0.720	1.27
OCH_3	1	1	3	0.12	0.602	1.960	1.649	1.67
CH_2OH	1	1	3	0.12	0.602	1.960	1.649	1.67
$NHCONH_2$	3	1	3	0.41	1.806	3.670	3.285	2.25
SO_2CH_3	3	1	3	1.89	0.977	4.270	1.302	6.26
SCH_3	1	1	3	0.29	0.602	2.350	2.039	2.06
$NHCH_3$	1	1	4	0.12	0.602	1.960	1.649	1.67
C_2H	0	2	1	–0.01	0.602	1.560	1.249	1.27
CH_2CN	1	2	2	0.00	1.431	2.490	2.243	1.31
C_2H_3	0	2	3	0.02	0.602	1.740	1.430	1.45
$COCH_3$	1	2	3	0.74	0.829	2.670	1.170	3.34
$COOCH_3$	2	2	3	0.88	2.408	3.630	1.118	3.73
$OCOCH_3$	2	2	3	0.39	1.806	3.630	3.248	2.21
CH_2COOH	2	2	3	0.39	1.806	3.630	3.248	2.21
OCH_2COOH	3	2	3	0.42	2.839	4.590	4.171	2.17
$NHCOCH_3$	2	2	4	0.39	1.802	3.630	3.248	2.21
$NHCOOCH_3$	3	2	4	0.76	3.495	4.590	2.381	3.30
C_2H_5	0	2	5	0.13	0.602	2.000	1.689	1.71
OC_2H_5	1	2	5	0.23	1.431	2.960	2.631	1.86
$N(CH_3)_2$	1	2	6	0.88	0.829	2.960	1.460	3.63
C_3H_5	0	3	5	0.55	0.829	2.571	1.500	2.81
$COOC_2H_5$	2	3	5	0.98	3.495	4.695	1.979	3.92
C_3H_7	0	3	7	0.25	1.431	3.000	2.667	1.90
$CH(CH_3)_2$	0	3	7	0.89	0.829	3.000	1.500	3.67
OC_3H_7	1	3	7	0.30	2.408	3.960	3.562	1.95
$OCH(CH_3)_2$	1	3	7	0.55	1.806	3.960	3.562	2.55
C_4H_9	0	4	9	0.32	2.408	4.000	3.600	1.99
$C(CH_3)_3$	0	4	9	1.74	0.976	4.000	1.111	5.91
OC_4H_9	1	4	9	0.32	3.495	4.960	4.533	1.89
NHC_4H_9	1	4	10	0.32	3.495	4.960	4.533	1.89
$N(C_2H_5)_2$	1	4	10	1.82	2.291	4.960	1.528	6.29
C_5H_{11}	0	5	11	0.34	3.495	5.000	4.571	1.93
C_6H_5	0	6	5	0.51	3.490	3.550	0.780	2.87
OC_6H_5	1	6	5	0.53	4.712	4.468	1.340	2.88
NHC_6H_5	1	6	6	0.53	4.712	4.468	1.340	2.88
C_6H_{11}	0	6	11	0.33	4.669	6.000	5.520	1.81
COC_6H_5	1	7	5	0.62	6.021	5.152	1.126	3.16
$CH_2CH_2C_6H_5$	0	8	9	0.59	6.021	5.471	1.926	3.00

studies are of great importance for molecular designers since they permit one to make proper choices of substituents for synthesis proposals and of parameters for QSAR evaluations.

We have previously reported some preliminary results of a pattern recognition study involving 59 common substituents and 58 descriptors [5]. These data are now extended to 74 descriptors and submitted to extensive analysis.

METHODS

The data set studied has $59 \times 74 = 4366$ values. Most of the values have been compiled from the literature. Unfortunately ca.18% of the data consist of missing values, all of which have been estimated and included in the data matrix [5, 11]. Estimates were made using a variety of techniques including regression equations from the original papers, linear regressions calculated from the present data set, and finally in some cases by a comparison with closely related substituents. Both data sets, with and without estimates, have been analysed. The descriptors used in this study are given in Table 1.

The following programs have been used: SPSS-X [57], CLUSTAN [58] and SIMCA-3B [59], running on either an IBM PC/AT, Norsk Data ND560 or CDC CYBER 170/855.

TABLE 3
LINEAR CORRELATIONS FOR $r > 0.80$ BETWEEN ADDITIONAL DESCRIPTORS (59–74) AND FULL DESCRIPTOR SET (1–74)

	B3	B4	LOLD	NA	NC	NHYDR	ESK	K0	K1	K3
VW		0.87		0.88						
PR				0.88						
B	0.83	0.81	0.83							
MR				0.88	0.913			0.86		
VTSAR			0.80	0.953	0.82	0.83		0.83	0.88	
VTSAL			0.82	0.957	0.86	0.86		0.84	0.917	
L			0.986					0.80		
B5			0.994							
SSAR	0.83	0.953								
SSAL	0.84	0.930								
ESC							0.84			0.83
X0AR								0.88	0.956	
X0VAR				0.88	0.84			0.85	0.85	
X1AR								0.939	0.917	
X1VAR				0.89	0.87			0.84	0.80	
SB									0.86	
MW									0.80	
B2	0.82									
B3										
NA					0.89	0.933				
NC				0.89						
ESK										
K1										

RESULTS

Linear regression analysis of parameters

A first approach to explore possible relationships among the 74 descriptors is the calculation of a correlation matrix. This was done by using the data matrix allowing for missing values. Part of this matrix has already been published [5]. Presently we restrict ourselves to significant correlations involving the additional descriptors (Nos. 59–74). The values of these descriptors are given in Table 2 which contains no estimated values. All of them are easily derived or calculated (see references from Table 1).

In Table 3 an abstract is given of the correlation matrix. The fact that 2 descriptors are highly correlated implies that they contain highly overlapping information content. It has been demonstrated by various authors that lipophilic parameters have at least 2 components. One of them is associated with bulk or steric properties, while the other(s) is (are) more electronic in nature and also related to hydrogen-bonding capability [5]. Therefore we have explored relationships of the type

$$\text{lipophilicity} = \text{bulk} + (\text{any other polarity-related parameter}) \tag{1}$$

Examples of successful correlations are the following:

$$\text{PIAR} = 0.052(\pm 0.002)\,\text{VTSAR} - 2.42(\pm 0.07)\,\text{K} - 0.65(\pm 0.06)$$

$$r = 0.992;\ n = 32;\ F = 961;\ s = 0.141 \tag{2}$$

$$\text{PIAR} = 0.74(\pm 0.09)\,\text{B5} - 2.53(\pm 0.24)\,\text{K} - 1.12(\pm 0.28)$$

$$r = 0.910;\ n = 32;\ F = 69.9;\ s = 0.481 \tag{3}$$

$$\text{FALR} = 0.042(\pm 0.004)\,\text{VTSAL} - 1.88(\pm 0.13)\,\text{HA} - 0.28(\pm 0.17)$$

$$r = 0.931;\ n = 52;\ F = 159;\ s = 0.460 \tag{4}$$

$$\text{FALHL} = -0.94(\pm 0.16)\,\text{ESC} + 1.09(\pm 0.11)\,\text{MUAL} + 0.87(\pm 0.21)$$

$$r = 0.938;\ n = 19;\ F = 58.6;\ s = 0.508 \tag{5}$$

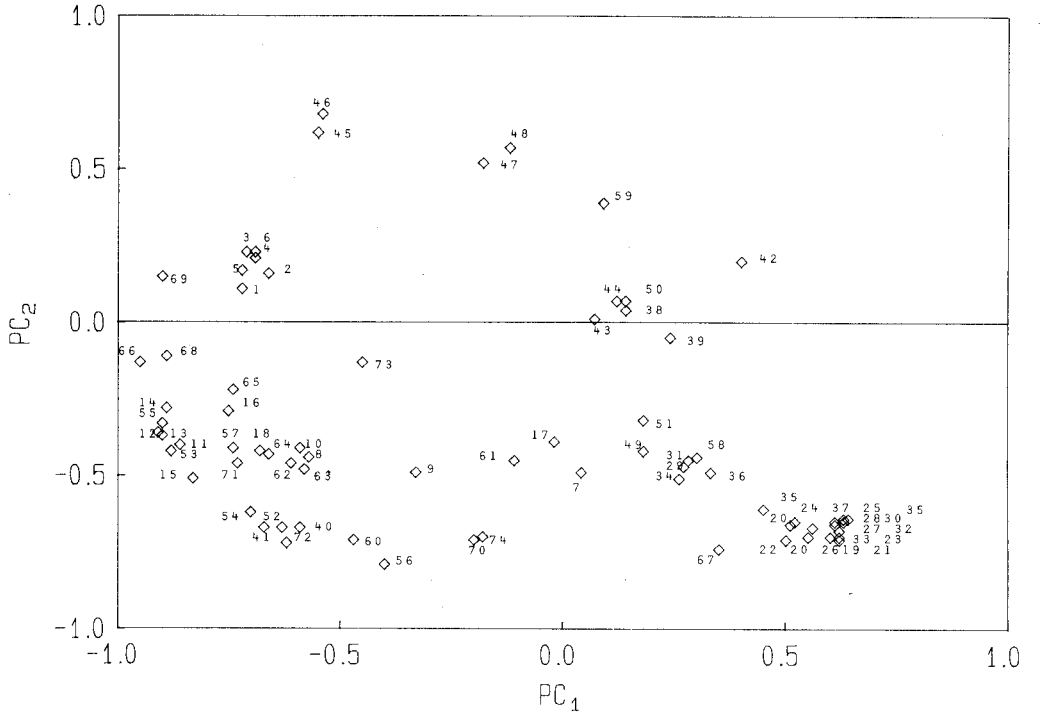
$$\text{FARR} = 0.081(\pm 0.008)\,\text{MR} - 0.51(\pm 0.04)\,\text{HB} + 0.16(\pm 0.14)$$

$$r = 0.931;\ n = 48;\ F = 145;\ s = 0.401 \tag{6}$$

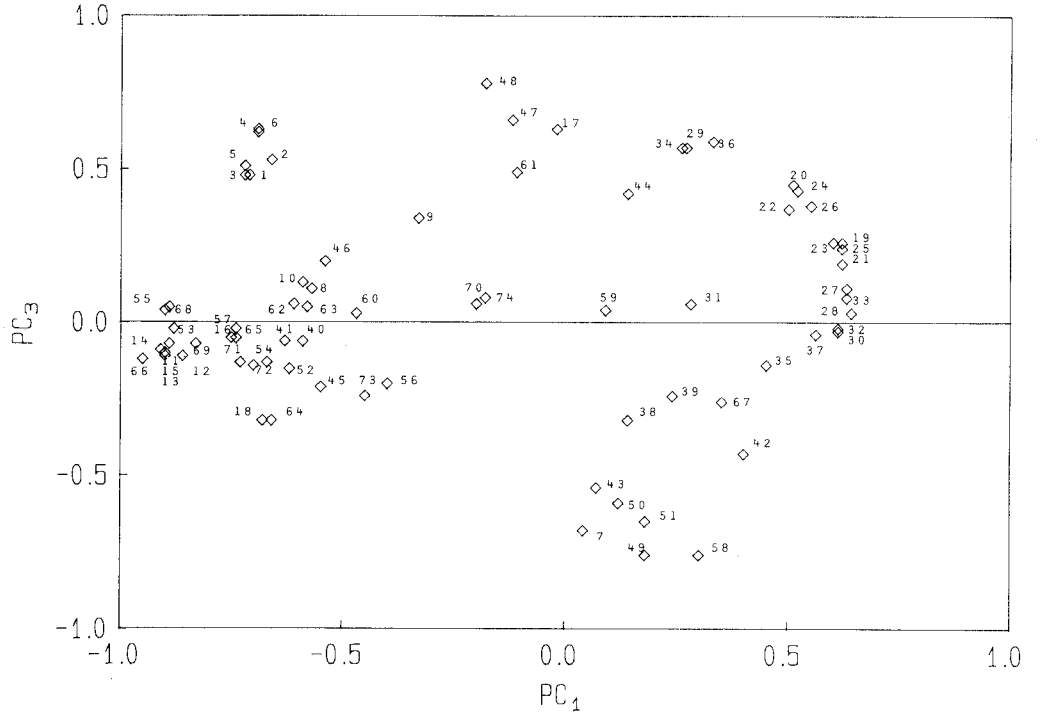
$$\text{FARHL} = 0.132(\pm 0.004)\,\text{MR} - 2.52(\pm 0.08)\,\text{K} - 0.24(\pm 0.06)$$

$$r = 0.992;\ n = 32;\ F = 848;\ s = 0.160 \tag{7}$$

A


1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
PC2
PC1

B


1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
PC3
PC1

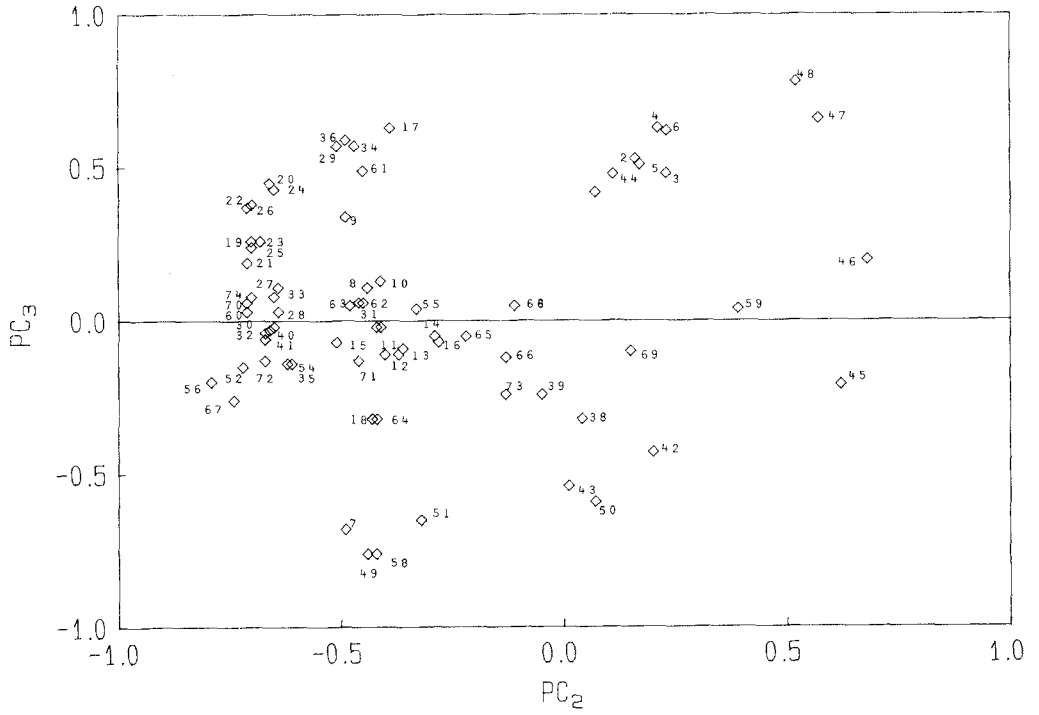


Fig. 1. Loading plots (parameters) of principal component analysis including all the 74 parameters.
A: PC_1 versus PC_2; B: PC_1 versus PC_3; C: PC_2 versus PC_3.

The standard errors are given in parentheses; r is the multiple regression coefficient, n the number of points (number of substituents), F the Fisher-test value for significance of the equation and s the standard deviation of the regression. In these correlations no values have been omitted and only those substituents are included for which literature values are available for the descriptors in the equations.

These equations confirm the bicomponent character of lipophilicity. The fragmental volume (VTSAR or VTSAL) or the molar refractivity (MR) accounts for the bulk term. The polarity term (including hydrogen-bonding capacity) is well-represented by HA, HB or K. The latter term is particularly interesting although it has received little attention in the literature.

Principal component analysis of parameters

Another technique for investigating the grouping of parameters and substituents in a multidimensional space is the reduction of this space to a small number of significant principal dimensions. This is done using principal component analyses [9]. Patterns in the data are of two kinds: relationships between variables (here property *parameters*) and relationships between objects (here *substituents*). The former are evaluated from a *loading plot* or *eigenvector plot*, the latter from a *scores plot*. Loadings are calculated from eigenvectors by multiplying each element by the square root of the corresponding eigenvalue λ.

TABLE 4
LOADINGS OF 5 SIGNIFICANT PRINCIPAL COMPONENTS

	Parameter	PC_1	PC_2	PC_3	PC_4	PC_5
1	PIAR	–0.72	0.11	0.48	0.42	–0.02
2	PIAL	–0.66	0.16	0.53	0.42	–0.03
3	FARR	–0.71	0.23	0.48	0.41	–0.02
4	FALR	–0.69	0.21	0.63	0.16	0.01
5	FARHL	–0.72	0.17	0.51	0.38	0.02
6	FALHL	–0.69	0.23	0.62	0.17	–0.04
7	K	0.04	–0.49	–0.68	–0.34	0.08
8	ES	–0.57	–0.44	0.11	–0.16	0.39
9	ESC	–0.33	–0.49	0.34	–0.17	0.44
10	UPS	–0.59	–0.41	0.13	–0.13	0.39
11	VW	–0.86	–040	–0.11	0.09	–0.11
12	VTSAR	–0.91	–0.36	–0.09	–0.00	–0.07
13	VTSAL	–0.90	–0.37	–0.11	0.00	–0.06
14	MR	–0.89	–0.28	–0.07	0.13	–0.09
15	PR	–0.83	–0.51	–0.07	0.02	–0.07
16	L	–0.75	–0.29	–0.05	0.09	–0.44
17	B1	–0.02	–0.39	0.63	–0.16	0.40
18	B5	–0.68	–0.42	–0.32	0.09	–0.18
19	SMHL	0.62	–0.68	0.26	0.22	–0.07
20	SPHL	0.51	–0.66	0.45	–0.07	–0.27
21	SMSW	0.62	–0.71	0.19	0.19	–0.02
22	SPSW	0.50	–0.71	0.37	–0.05	–0.27
23	SMM	0.60	–0.70	0.26	0.21	–0.07
24	SPM	0.52	–0.65	0.43	–0.13	–0.27
25	SME	0.62	–0.70	0.24	0.19	–0.03
26	SPE	0.55	–0.70	0.38	–0.04	–0.17
27	SSTAR	0.63	–0.64	0.11	0.36	0.00
28	SIM	0.64	–0.64	0.03	0.36	0.06
29	SRM	0.27	–0.47	0.57	–0.41	–0.40
30	SICH	0.61	–0.65	–0.02	0.39	0.09
31	SRCH	0.28	–0.45	0.06	–0.42	–0.38
32	SIBR	0.61	–0.66	–0.03	0.39	0.09
33	F	0.63	–0.65	0.08	0.37	0.06
34	R	0.26	–0.51	0.57	–0.37	–0.42
35	FNEW	0.45	–0.61	–0.14	0.51	0.15
36	RNEW	0.33	–0.49	0.59	–0.30	–0.39
37	SF	0.56	–0.67	–0.04	0.38	0.12
38	SX	0.14	0.04	–0.32	0.60	0.12
39	IOTA	0.24	–0.05	–0.24	0.58	0.03
40	SSAL	–0.59	–0.67	–0.06	–0.02	0.17
41	SSAR	–0.63	–0.67	–0.06	–0.07	0.13
42	E	0.40	0.20	–0.43	0.47	0.01
43	RE	0.07	0.01	–0.54	0.65	–0.10
44	I	0.14	0.07	0.42	–0.54	0.09
45	MUAR	–0.55	0.62	–0.21	–0.13	0.00
46	MUAL	–0.54	0.68	0.20	–0.16	–0.05
47	LAMDAR	–0.12	0.57	0.66	0.40	0.04

TABLE 4 continued
LOADINGS OF 5 SIGNIFICANT PRINCIPAL COMPONENTS

	Parameter	PC_1	PC_2	PC_3	PC_4	PC_5
48	LAMDAL	–0.18	0.52	0.78	0.13	0.05
49	HA	0.18	–0.42	–0.76	–0.08	–0.07
50	HD	0.12	0.07	–0.59	–0.49	–0.10
51	HB	0.18	–0.32	–0.65	–0.47	–0.13
52	X0AR	–0.62	–0.72	–0.15	–0.03	–0.05
53	X0VAR	–0.88	–0.42	–0.02	0.13	–0.00
54	X1AR	–0.70	–0.62	–0.14	–0.00	–0.15
55	X1VAR	–0.90	–0.33	0.04	0.13	–0.06
56	SB	–0.40	–0.79	–0.20	–0.13	0.12
57	B	–0.74	–0.41	–0.02	0.00	–0.12
58	C	0.30	–0.44	–0.76	–0.10	–0.08
59	RAND	0.09	0.39	0.04	–0.05	0.08
60	MW	–0.47	–0.71	0.03	0.16	0.03
61	B1OLD	–0.11	–0.45	0.49	–0.28	0.40
62	B2	–0.61	–0.46	0.06	–0.03	0.35
63	B3	–0.58	–0.48	0.05	–0.10	0.36
64	B4	–0.66	–0.43	–0.32	0.08	–0.19
65	LOLD	–0.74	–0.22	–0.05	0.09	–0.45
66	NA	–0.95	–0.13	–0.12	–0.03	–0.07
67	NMET	0.35	–0.74	–0.26	–0.11	0.05
68	NC	–0.89	–0.11	0.05	0.09	–0.15
69	NHYDR	–0.90	0.15	–0.10	–0.04	–0.03
70	ESK	–0.20	–0.71	0.06	–0.24	0.50
71	K0	–0.73	–0.46	–0.13	0.11	–0.28
72	K1	–0.67	–0.67	–0.13	–0.02	–0.06
73	K3	–0.45	–0.13	–0.24	0.14	–0.46
74	KSI	–0.18	–0.70	0.08	–0.25	0.53

The full data set (including the estimated values) was analysed with the SIMCA program. On the basis of cross-validation [60], a statistical method to evaluate the significance of the principal component, 5 significant principal components emerged. These principal components explained 83.94% of data variance as follows: $PC_1 = 33.46\%$, $PC_2 = 25.07\%$, $PC_3 = 13.12\%$, $PC_4 = 7.48\%$, $PC_5 = 4.82\%$. The eigenvalues were as follows: $\lambda_1 = 24.76$, $\lambda_2 = 18.55$, $\lambda_3 = 9.71$, $\lambda_4 = 5.53$ and $\lambda_5 = 3.57$. A number of loading plots were examined and the most informative ones are presented in Fig. 1. Some interesting conclusions emerge. As expected, the parameters cluster into more or less well defined groups. Clearly the lipophilic descriptors 1–6 form a subclass. The steric descriptors are seen in the lower left corner, and the electronic descriptors in the lower right corner of Fig. 1A. A number of less well defined descriptors are found around the (0,0) point. Indeed, it is known that variables with little information in the *considered plane* are found in this area [60].

Surprisingly, in a preliminary analysis (results not shown), the steric parameters ES (8) and ESC (9) are found in the upper right corner, far removed from the other steric parameters. This can be explained by the fact that all parameter values were taken as tabulated in the literature without a priori questioning the sign. Closer inspection indicates that the sign of these 2 parame-

ters must be inverted to bring them in the region of the other steric parameters. This was done in Fig. 1. Another satisfactory aspect is the fact that random numbers (RAND, 59) project close to the origin (0,0). This was confirmed by further trials with other random numbers showing that RAND oscillates around (0,0) and is, as expected, devoid of information.

The loadings of the 5 significant PCs are given in Table 4. In this table, a loading value close to unity for a given parameter reflects a significant contribution to a given principal component. A visual representation of these contributions for some representative parameters is given in Fig. 2; for example, the steric parameter MR (14) significantly contributes to PC_1 and poorly contributes to the other principal components while the hydrogen-bonding parameter HA (49) contributes mainly to PC_3. Poor contribution of a given parameter is arbitrarily defined for a loading value < 0.4.

Cluster analysis of substituents

Although not always easy to use in practice, cluster analysis has become, besides principal component analysis, another important tool in chemometrics. Indeed, objects (here substituents) in multidimensional space can be agglomerated in various ways using different hierarchical clustering methods [10]. The basic procedure with all these methods is similar. Briefly, they start with the computation of a similarity or distance matrix between the objects and end with a dendrogram showing the successive fusions of objects until a stage where all the objects are in one group. Differences between methods are related to the different ways of defining distance or similarity between an object and a group containing several objects, or between groups of objects. Among the hierarchical clustering techniques, Ward's method is considered as the best of the hierarchical options. In this study, Ward's method was chosen to classify the 59 substituents using the 74 parameters. This method provided an informative dendrogram (Fig. 3) in which 6 clusters can be derived as reported in Table 5. Schematically, these clusters contain the following substituents:

A: halogens and some small, strongly electron-withdrawing substituents;
B: lower alkyls, small amino and alkoxy groups;
C: sulfo-containing groups, carbonyl and ester groups;
D: carboxyl, carboxamido and analogues groups;
E: higher alkyls;
F: larger amino and alkoxy groups.

Globally however, the classification is not fully satisfactory for a chemist. In addition, other CA techniques yielded somewhat different classifications (not shown). We therefore conclude that CA falls short of extracting much useful information from the data.

Principal component analysis of substituents

Very informative are the scores plots in which the grouping of the substituents with similar properties can be evaluated. Using the scores values of the two first principal components t_1 and t_2, respectively, and a physicochemical hypothesis, an interesting grouping of the substituents was seen (Fig. 4). In this figure, the substituents are classified into 5 groups, four of which define 4 parallel lines. Five other substituents form an additional group. In Table 6 we have collected the groups defined in Fig. 4. Group I comprises neutral aliphatic groups and *N*-alkyl substituents. In

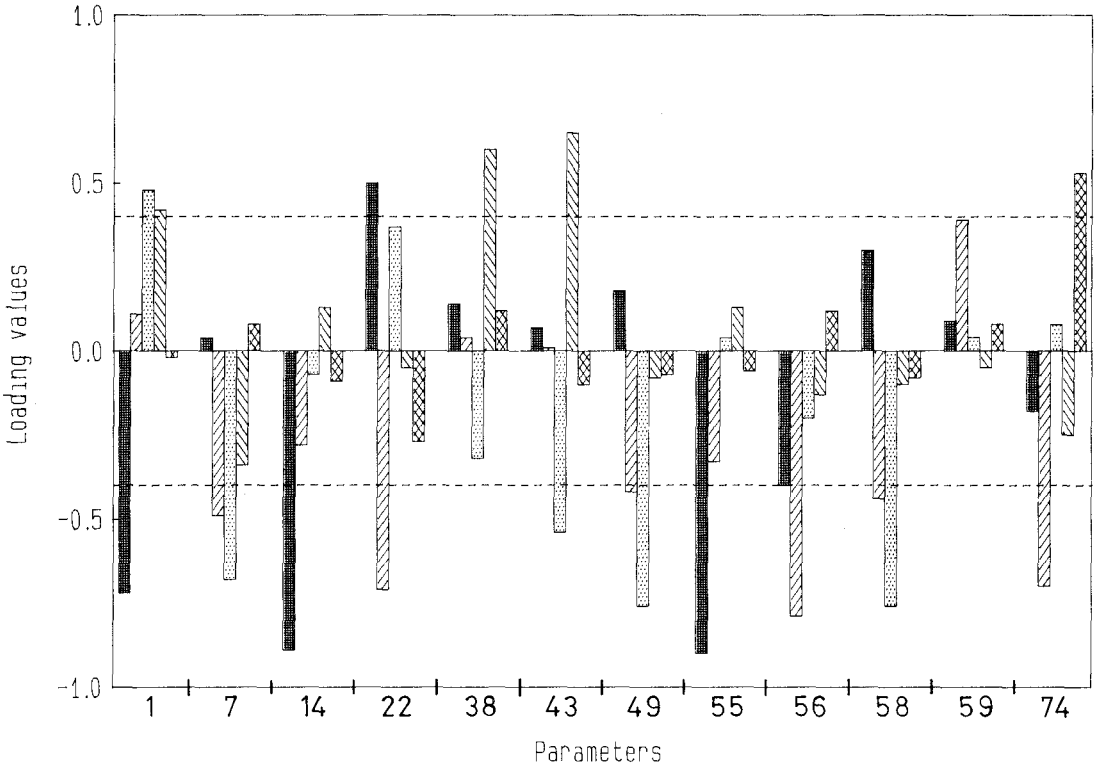


Fig. 2. Contribution of some representative parameters (1, 7, 14, 22, 38, 43, 49, 55, 56, 58, 59, 74) to the 5 principal components as a function of their loading values (bar plot; values close to unity imply a high contribution, values smaller than 0.4 are taken to mean poor contribution); symbols are PC_1, PC_2, PC_3, PC_4 and PC_5.

group II the character becomes somewhat more polar and all *O*-alkyl substituents are found. In groups III–V substituents of increasing polarity are seen. Within each group, there is a global increase in the bulk factor when going from left to right along the parallel lines, and we hypothesize that orthogonally to these parallels the factor polarity is operating, as visualized in Fig. 5. In similar studies in the literature [9] such plots have revealed groups for, for example, alkyls, halogens, polar substituents.

The above grouping of substituents is based on a t_1 versus t_2 plot and on physicochemical arguments. However, such a classification can be fully satisfactory only when it is based on statistical approaches such as SIMCA, K nearest neighbor, linear discriminant analysis or related methods. All these classification methods are means to mathematically describe the position of a given group in a multidimensional space. In this study the SIMCA method was used and it confirmed the classification in Table 6. When the SIMCA classification was carried out with a smaller set of substituents, those not included were later classified correctly. For more detail the reader is referred to the excellent paper of Dunn and Wold [61].

TABLE 5
SUBSTITUENT CLASSIFICATION FROM WARD'S METHOD OF CLUSTERING (FIG. 3)*

Group A	Group B	Group C	Group D	Group E	Group F
1 BR	6 H	10 SO_2NH_2	19 COOH	42 C_3H_5	46 OC_3H_7
2 Cl	22 CH_3	26 SO_2CH_3	20 $CONH_2$	44 C_3H_7	50 OC_4H_9
4 I	31 C_2H_3	13 SO_2CF_3	24 CH_2OH	45 $CH(CH_3)_2$	47 $OCH(CH_3)_2$
8 SH	39 C_2H_5	11 CF_3	30 CH_2CN	49 $C(CH_3)_3$	51 NHC_4H_9
29 CCH	7 OH	32 $COCH_3$	21 $OCONH_2$	48 C_4H_9	52 $N(C_2H_5)_2$
27 SCH_3	9 NH_2	33 $COOCH_3$	36 OCH_2COOH	53 C_5H_{11}	55 OC_6H_5
3 F	28 $NHCH_3$	43 $COOC_2H_5$	35 CH_2COOH	54 C_6H_5	56 NHC_6H_5
5 NO_2	23 OCH_3	12 OCF_3	25 $NHCONH_2$	57 C_6H_{11}	58 COC_6H_5
15 CN	40 OC_2H_5	14 SCF_3	37 $NHCOCH_3$	59 $CH_2CH_2C_6H_5$	
18 CHO	41 $N(CH_3)_2$	34 $OCOCH_3$	38 $NHCOOCH_3$		
		16 SCN			
		17 NCS			

*Ordering as in dendrogram.

TABLE 6
SUBSTITUENT CLASSIFICATION BASED ON SCORES PLOT t_1-t_2 (FIG. 4)*

Group I	Group II	Group III	Group IV	Group V
6 H	7 OH	3 F	18 CHO	15 CN
9 NH_2	8 SH	2 Cl	19 COOH	5 NO_2
22 CH_3	24 CH_2OH	1 Br	20 $CONH_2$	10 SO_2NH_2
28 $NHCH_3$	29 CCH	4 I	16 SCN	26 SO_2CH_3
31 $CHCH_2$	23 OCH_3	30 CH_2CN	32 $COCH_3$	13 SO_2CF_3
39 C_2H_5	27 SCH_3	21 $OCONH_2$	17 NCS	
41 $N(CH_3)_2$	40 OC_2H_5	37 $NHCOCH_3$	12 OCF_3	
42 C_3H_5	47 $OCH(CH_3)_2$	35 CH_2COOH	11 CF_3	
44 C_3H_7	46 OC_3H_7	36 OCH_2COOH	34 $OCOCH_3$	
45 $CH(CH_3)_3$	50 OC_4H_9	25 $NHCONH_2$	33 $COOCH_3$	
48 C_4H_9	54 C_6H_5	38 $NHCOOCH_3$	14 SCF_3	
49 $C(CH_3)_3$	56 NHC_6H_5	55 OC_6H_5	43 $COOC_2H_5$	
51 NHC_4H_9			58 COC_6H_5	
52 $N(C_2H_5)_2$				
53 C_5H_{11}				
57 C_6H_{11}				
59 $CH_2CH_2C_6H_5$				

* The substituents are ordered according to increasing t_1.

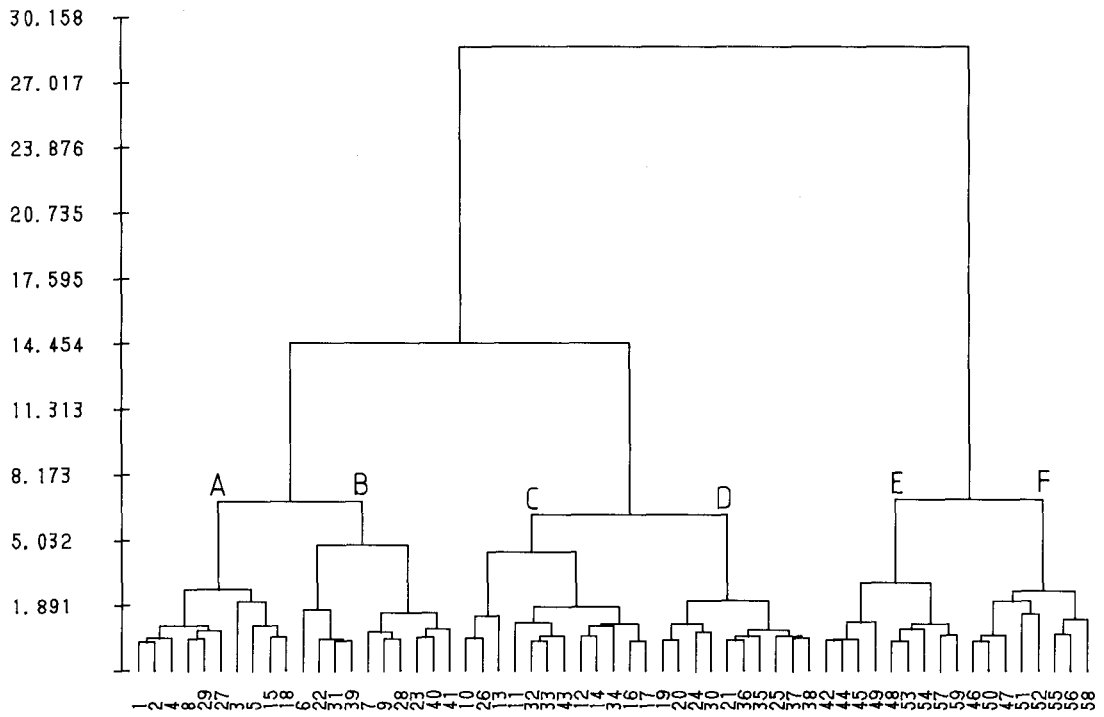


Fig. 3. Dendrogram obtained from hierarchical clustering of 59 substituents and based on 74 parameters, Ward's method. Six clusters are indicated by the symbols A, B, C, D, E, F and reported in Table 5.

DISCUSSION

From a data set of 74 parameters describing properties of 59 substituents, the present statistical study extracts correlative information of interest in physical chemistry and drug design. In particular, the database we have compiled here and elsewhere [5] is of significance to molecular designers since it permits a rational choice among substituents for synthesis planning, and among parameters for QSAR studies. Studies of this type are not new, but the size of the matrix (4366 data values) makes the present investigation particularly comprehensive and its results robust. However, it can be noted that the extension of our data set from 58 descriptors [5] to 74 descriptors did not significantly influence the results.

Correlations between, and relationships among, parameters are explored in Eq. 2–7, Table 3, and Fig. 1. Redundancies within groups of QSAR parameters are thus made explicit. Exploring the loading values of parameters reveals that some of them have a low information content in the first two principal components. This, however, does not imply uselessness since some of such parameters (e.g., K, HA, HB) appear to contain 'second degree information' which becomes apparent in Eq. 2–7. This must be linked to the fact that they have significant loadings in PC_3, PC_4 or PC_5.

For the 20 natural amino acids Wold and collaborators have derived a new set of parameters [62, 63] which they called principal properties and designated z_1, z_2 and z_3. These new parameters

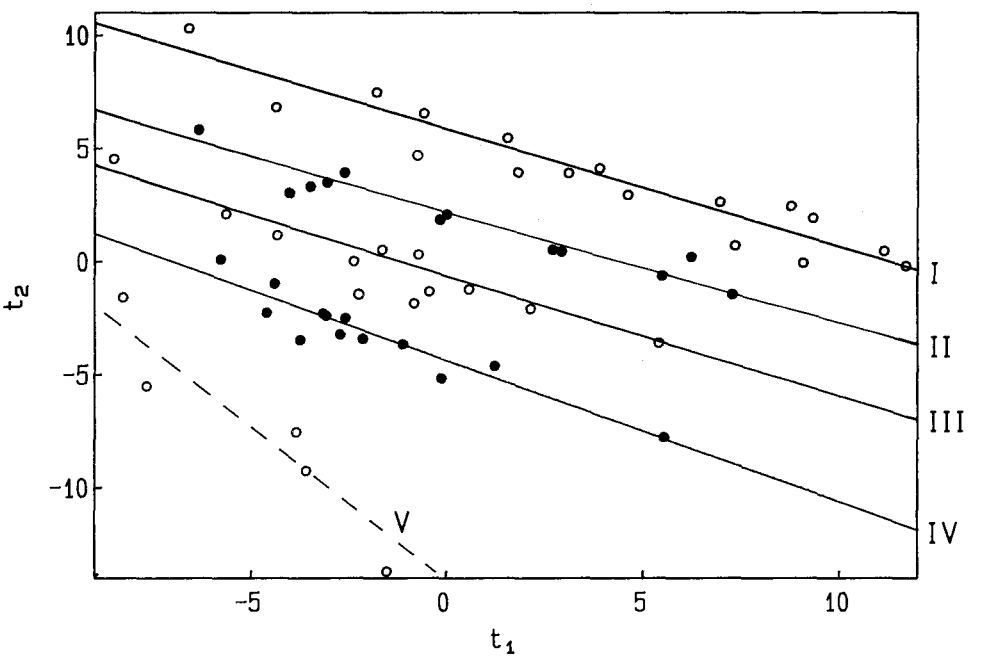


Fig. 4. Scores plot of principal components for 59 substituents from 74 parameters. The classification into 5 groups was confirmed by the SIMCA method and is given in Table 6. Lines I - IV were calculated by linear regression.

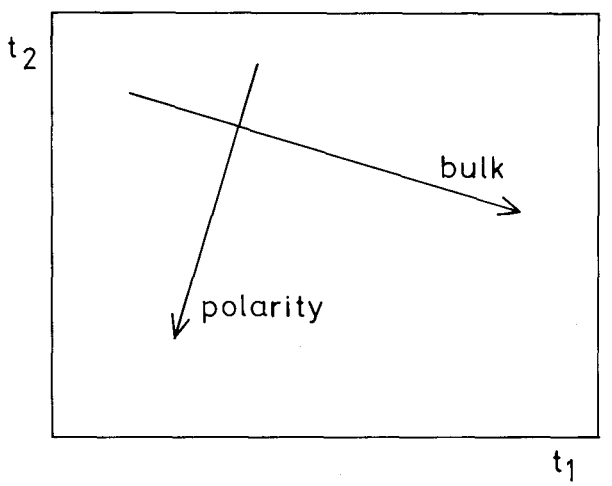


Fig. 5. Information content extracted from scores plot t_1 versus t_2.

were obtained from the vector scores of 3 latent variables in a PCA and tentatively interpreted as hydrophilicity, bulk and electronic effect, respectively. Such new principal properties are potentially useful as design variables in a multivariate design approach. Earlier Cramer [53] had derived

TABLE 7
PRINCIPAL PROPERTIES DERIVED FROM 74 DESCRIPTORS

Substituent	W_1	W_2	W_3	W_4	W_5
Br	–4.32	1.07	–3.28	2.85	1.32
Cl	–5.63	2.01	–2.85	2.88	0.84
F	–8.50	4.45	–0.45	5.44	0.90
I	–2.37	–0.09	–3.80	2.74	1.62
NO_2	–7.66	–5.59	–2.03	1.80	0.58
H	–6.59	10.30	–2.62	–2.11	–2.55
OH	–6.32	5.74	3.94	1.46	1.38
SH	–4.02	2.92	–1.69	0.52	0.25
NH_2	–4.36	6.81	4.93	–1.07	1.97
SO_2NH_2	–3.85	–7.64	1.31	–5.18	1.37
CF_3	–2.71	–3.32	–5.71	–0.74	3.06
OCF_3	–2.14	–3.52	0.51	1.57	–0.57
SO_2CF_3	–1.53	–13.80	–1.90	–0.24	2.58
SCF_3	–0.12	–5.27	–2.21	1.69	–0.75
CN	–8.26	–1.67	–1.70	1.57	–2.01
SCN	–3.75	–3.56	–1.07	2.43	–1.33
NCS	–3.15	–2.39	–1.07	4.46	–1.72
CHO	–5.77	0.01	–1.46	–1.62	–1.09
COOH	–4.38	–1.07	–0.52	–2.73	–0.91
$CONH_2$	–4.59	–2.37	1.81	–4.32	–1.45
$OCONH_2$	–2.25	–1.60	3.64	–2.38	–2.63
CH_3	–1.79	7.48	–2.99	–2.21	0.51
OCH_3	–3.05	3.39	3.09	3.17	0.80
CH_2OH	–2.60	3.82	1.00	–4.48	–0.94
$NHCONH_2$	–0.72	0.20	5.32	–2.72	0.10
SO_2CH_3	–3.59	–9.36	0.27	–1.69	1.79
SCH_3	–0.16	1.73	–1.37	–0.00	0.72
$NHCH_3$	–0.59	6.54	5.56	–0.56	1.93
C_2H	–3.49	3.22	–3.23	0.24	–0.92
CH_2N	–1.65	0.42	0.76	–1.37	0.37
C_2H_3	–0.77	4.62	–3.17	–0.80	–0.58
$COCH_3$	–3.07	–2.50	–1.33	–2.12	–0.62
$COOCH_3$	–1.12	–3.76	–1.24	–1.91	–0.26
$OCOCH_3$	–2.58	–2.60	3.53	2.39	–1.86
CH_2COOH	–0.43	–1.43	1.54	–3.11	–1.25
OCH_2COOH	0.58	–1.34	4.19	–2.10	–2.67
$NHCOCH_3$	–0.83	–1.96	4.77	–0.65	–0.54
$NHCOOCH_3$	2.14	–2.21	3.78	–1.40	0.16
C_2H_5	1.55	5.45	–2.94	–1.73	–0.34
OC_2H_5	0.02	1.98	2.42	2.08	–0.63
$N(CH_3)_2$	1.81	3.93	4.65	1.13	3.77
C_3H_5	3.10	3.90	–3.23	–1.28	2.28
$COOC_2H_5$	1.22	–4.73	–1.18	–0.83	–1.60
C_3H_7	3.89	4.12	–2.99	–0.49	–0.91
$CH(CH_3)_2$	4.60	2.94	–3.98	–2.08	2.30
OC_3H_7	2.70	0.40	2.62	3.06	–1.38
$OCH(CH_3)_2$	2.93	0.32	3.35	3.29	1.71

TABLE 7 continued
PRINCIPAL PROPERTIES DERIVED FROM 74 DESCRIPTORS

Substituent	W_1	W_2	W_3	W_4	W_5
C_4H_9	6.95	2.62	–2.57	0.24	–2.01
$C(CH_3)_3$	7.34	0.69	–5.23	–2.76	4.85
OC_4H_9	5.49	–0.76	3.00	4.01	–2.12
NHC_4H_9	8.76	2.45	3.88	–2.09	–2.96
$N(C_2H_5)_2$	9.07	–0.07	4.40	0.11	5.80
C_5H_{11}	9.33	1.92	–2.71	0.70	–3.26
C_6H_5	6.23	0.08	–3.43	0.55	–0.55
OC_6H_5	5.40	–3.72	1.89	3.92	0.15
NHC_6H_5	7.28	–1.58	4.70	0.83	1.16
C_6H_{11}	11.14	0.45	–3.66	0.31	–0.82
COC_6H_5	5.53	–7.89	–0.47	0.09	–0.91
$CH_2CH_2C_6H_5$	11.69	–0.20	–2.78	1.22	–2.11

TABLE 8
CORRELATION COEFFICIENTS BETWEEN PRINCIPAL PROPERTIES AND ORIGINAL DESCRIPTORS

No.	Descriptor	W_1	W_2	W_3
1	PIAR	.703	.08	−.56
2	PIAL	.57	.05	−.50
3	FARR	.707	.22	−.53
4	FALR	.687	.13	−.636
5	FARHL	.719	.15	−.56
6	FALHL	.682	.17	−.671
7	K	.04	−.43	.750
8	ES	−.616	.53	.42
9	ESC	−.31	.753	.45
10	UPS	.59	−.33	−.10
11	VW	.837	−.51	.25
12	VTSAR	.908	−.37	.08
13	VTSAL	.936	−.37	.07
14	MR	.896	−.29	.02
15	PR	.826	−.52	.04
16	L	.758	−.30	−.06
17	B1	−.01	−.36	−.50
18	B5	.688	−.44	.27
19	SMHL	−.644	−.651	−.29
20	SPHL	−.53	−.626	−.511
21	SMSW	−.53	−.778	−.13
22	SPSW	−.41	−.777	−.27
23	SMM	−.56	−.766	−.19
24	SPM	−.48	−.713	−.41
25	SME	−.61	−.718	−.16
26	SPE	−.53	−.721	−.32

TABLE 8 continued
CORRELATION COEFFICIENTS BETWEEN PRINCIPAL PROPERTIES AND ORIGINAL DESCRIPTORS

No.	Descriptor	W_1	W_2	W_3
27	SSTAR	−.657	−.632	.00
28	SIM	−.603	−.683	.04
29	SRM	−.20	−.57	−.58
30	SICH	−.58	−.699	.14
31	SRCH	−.30	−.45	−.642
32	SIBR	−.58	−.669	.14
33	F	−.624	−.665	−.08
34	R	−.32	−.52	−.607
35	FNEW	−.23	−.56	.33
36	RNEW	−.43	−.48	−.645
37	SF	−.58	−.700	.05
38	SX	−.26	.02	.40
39	IOTA	−.45	−.09	.23
40	SSAL	.617	−.677	.04
41	SSAR	.632	−.654	−.01
42	E	−.40	.20	.31
43	RE	.00	.01	.40
44	I	−.19	.07	−.32
45	MUAR	.48	.59	.18
46	MUAL	.635	.630	−.31
47	LAMDAR	.11	.56	−.697
48	LAMDAL	.10	.45	−.790
49	HA	−.16	−.41	.763
50	HD	−.10	.10	.59
51	HB	−.14	−.26	.661
52	X0AR	.623	−.731	.14
53	X0VAR	.880	−.42	−.00
54	X1AR	.701	−.639	.11
55	X1VAR	.894	−.34	−.08
56	SB	.41	−.791	.25
57	B	.685	−.50	−.06
58	C	−.57	−.25	.738
59	RAND	−.09	.36	−.04
60	MW	.47	−.690	−.04
61	B1OLD	.12	−.43	−.33
62	B2	.60	−.45	−.04
63	B3	.55	−.48	−.04
64	B4	.657	−.44	.24
65	LOLD	.761	−.22	−.10
66	NA	.953	−.15	.08
67	NHET	−.34	−.719	.28
68	NC	.892	.14	.10
69	NHYDR	.908	.15	.07
70	ESK	.20	−.699	.07
71	K0	.740	−.48	.05
72	K1	.670	−.681	.11
73	K3	.46	−.15	.13
74	KSI	.17	−.685	.05

chemical descriptor scales BC(DEF) for organic compounds by a similar technique.

From the present data set we have calculated the vector scores for the substituents and defined a set of substituent principal properties W_1, W_2, W_3, W_4 and W_5 which are given in Table 7. The correlation between some of these scores (W_1, W_2 and W_3) with the original parameters (Table 8) reveals that W_1 is correlated ($r > 0.6$) with VW, PR, MR, VTSAR, VTSAL, X0VAR, X1VAR, NA, NC and NHYDR. W_2 shows a correlation with SMSW, SPSW, SMM, SPM, SME, SPE, ESC, X0AR, SB and NHET, while W_3 correlates with HA, HB, C, K, LAMDAR and LAMDAL. This approximately means that W_1 encodes molecular bulk, W_2 describes electronic character, and W_3 reflects hydrogen bonding capability.

Analogies among substituents are explored in Tables 5–7 and Fig. 3 and 4. The 59 substituents were classified into 6 major groups using Ward's method of cluster analysis, but other CA methods yielded somewhat different results. An alternative classification based on a physicochemical hypothesis (Fig. 4 and Table 6) and verified by the SIMCA method seems more informative since it assembles substituents of increasing bulk into 5 groups of increasing polarity. It was mentioned earlier in this paper that lipophilicity can be seen as the sum of bulk and polarity and, interestingly, the 2 factors emerge as vectors in Fig. 5. Correlating lipophilicity with the scores t_1, t_2 and perhaps t_3 could allow its interpretation in vector terminology, i.e.,

$$\text{lipophilicity} = \text{bulk} + \text{polarity} \tag{8}$$

To do so, however, the 4 parallel lines in Fig. 4 should be given nul slope by rotating the system of axes. The highly interesting properties and possible applications of these rotated scores will be discussed in a forthcoming paper (in preparation).

ACKNOWLEDGEMENT

The authors thank the Swiss National Science Foundation for grant 3.508-0.86.

REFERENCES

1 Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616–1626.
2 Hansch, C. and Leo, A., Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, 1979.
3 Hansch, C., Unger, S.H. and Forsythe, A.B., J. Med. Chem., 16 (1973) 1217–1222.
4 Leo, A., Hansch, C. and Elkings, D., Chem. Rev., 71 (1971) 525–616.
5 Van de Waterbeemd, H. and Testa, B., In Testa, B. (Ed.) Advances in Drug Research, Vol. 16, Academic Press, London, 1987, pp. 85–225.
6 Franke, R., Theoretical Drug Design Methods, Elsevier, Amsterdam, 1984.
7 Tichy, M., Int. J. Quantum Chem., 16 (1979) 509–515.
8 Dearden, J.C. and Mays, P.K., J. Pharm. Pharmacol., 37 (1985) 70P.
9 Alluni, S., Clementi, S., Edlund, U., Johnels, D., Hellberg, S., Sjöström, M. and Wold, S., Acta Chem. Scand., B37 (1983) 47–53.
10 Everitt, B., Cluster Analysis, Halsted Press, New York, 1980.
11 Van de Waterbeemd, H. and Carrupt, P.A., DESBASE: A Multiparameter Substituent Database, Softarts-Actimol, Lausanne, 1988.
12 Hansch, C. and Leo, A., Pomona College Med. Chem. Project log P Data Bank, 1983.
13 Martin, Y.C., Quantitative Drug Design: A Critical Introduction, Dekker, New York, 1978.
14 Rekker, R.F. and De Kort, H.M., Eur. J. Med. Chem., 14 (1979) 479–488.

15 Van de Waterbeemd, H., Hydrophobicity of Organic Compounds, Booksoft Vol. 1, Darvas, F. (Ed.), Compudrug, Budapest, 1986.
16 Livingstone, D.J., Hyde, R.M. and Foster, R., Eur. J. Med. Chem., 14 (1979) 393–397.
17 Exner, O., In Chapman, N.B. and Shorter, J. (Eds.) Correlation Analysis in Chemistry, Plenum Press, New York, 1978, pp. 439–540.
18 Taft, R.W., In Newman, M.S. (Ed.) Steric Effects in Organic Chemistry, Wiley, New York, 1956, pp. 556.
19 Dunn, W.J., Eur. J. Med. Chem., 12 (1977) 109–112.
20 Hancock, K., Meyers, E.A. and Yager, B.J., J. Am. Chem. Soc., 83 (1961) 4211–4242.
21 Charton, M., In Roche, E.B. (Ed.) Design of Biopharmaceutical Properties through Prodrugs and Analogs, Am. Pharm. Ass., Washington, DC, 1977, pp. 228–281.
22 Charton, M., In Charton, M. and Motoc, I. (Eds.) Steric Effects in Drug Design, Springer-Verlag, Berlin, 1983, pp. 57–91.
23 Bondi, A., J. Phys. Chem., 68 (1964) 441–451.
24 Moriguchi, I., Kanada, Y. and Komatsu, K., Chem. Pharm. Bull., 24 (1976) 1799–1806.
25 Moriguchi, I. and Kanada, Y., Chem., Pharm. Bull., 25 (1977) 926–935.
26 Yang, G., Lien, E.J. and Guo, Z., Quant. Struct.-Act. Relatsh., 5 (1986) 12–18.
27 Testa, B. and Seiler, P., Arzneim.-Forsch., 31 (1981) 1053–1058.
28 Mager, P.P., Multidimensional Pharmacochemistry: Design of Safer Drugs, Academic Press, Orlando, 1984.
29 Norrington, F.E., Hyde, R.M., Williams, S.G. and Wootton, R., Eur. J. Med. Chem., 18 (1975) 604–607.
30 Ahmad, P., Fyfe, C.A. and Mellors, A., Biochem. Pharmacol., 24 (1975) 1103–1109.
31 Ahmad, P., Fyfe, C.A. and Mellors, A., Can. J. Biochem., 53 (1975) 1047–1053.
32 Exner, O., Collect. Czech. Chem. Commun., 32 (1967) 1–23, 24–54.
33 Quayle, O.R., Chem. Rev., 53 (1953) 439–589.
34 Verloop, A., In Miyamato, J. (Ed.) IUPAC Pesticide Chemistry, Pergamon, Oxford, 1983, pp. 339–344.
35 Verloop, A., The STERIMOL Approach to Drug Design, in preparation.
36 Sjöström, M. and Wold, S., Chem. Script., 9 (1976) 200–210.
37 Mager, P.P., Sci. Pharm., 48 (1980) 117–126.
38 Hine, J., Structural Effects on Equilibrium in Organic Chemistry, Wiley, New York, 1975, pp. 55–102.
39 Li, W.Y., Guo, Z.R. and Lien, E.J., J. Pharm. Sci., 73 (1984) 553–558.
40 Charton, M., Prog. Phys. Org. Chem., 13 (1981) 119–251.
41 Bijloo, G.J. and Rekker, R.F., Quant, Struct.-Act. Relatsh. 3 (1984) 91–96, 111–115.
42 Swain, C.G. and Lupton, E.C., J. Am. Chem. Soc., 90 (1968) 4328–4337.
43 Swain, C.G., Unger, S.H., Rosenquist, N.R. and Swain, M.S., J. Am. Chem. Soc., 105 (1983) 492–502.
44 Marriott, S. and Topsom, R.D., Tetrahedron Lett., 23 (1982) 1485–1488.
45 Marriott, S., Reynolds, W.F., Taft, R.W. and Topsom, R.D., J. Org. Chem., 49 (1984) 959–965.
46 Inamoto, N. and Masuda, S., Tetrahedron Lett., 18 (1977) 3287–3290.
47 Inamoto, N., Masuda, S., Tori, K. and Yoshimura, Y., Tetrahedron Lett., 19 (1978) 4547–4550.
48 Sasaki, Y., Takagi, T., Yamazato, Y., Iwata, A. and Kawaki, H., Chem. Pharm. Bull., 29 (1981) 3073–3075.
49 Esaki, T., J. Pharmacobiodyn., 3 (1980) 562–576.
50 Lien, E.J., Guo, Z.R., Li, R.L. and Su, C.T., J. Pharm. Sci., 71 (1982) 641–655.
51 Kier, L.B. and Hall, L.H., J. Pharm. Sci., 70 (1981) 583–589.
52 Austel, V., Kutter, E. and Kalbfleisch, W., Arzneim.-Forsch., 29 (1979) 585–587.
53 Cramer, R.D., J. Am. Chem. Soc., 102 (1980) 1837–1849, 1849–1859.
54 Verloop, A., Hoogenstraaten, W. and Tipker, J., In Ariëns, E.J, (Ed.) Drug Design, Vol. VII, Academic Press, New York, 1976, pp. 165-207.
55 Kier, L.B., Quant. Struct.-Act. Relatsh., 5 (1986) 1–7, 7–12.
56 Kier, L.B., Med. Chem. Revs., 7 (1987) 417–440.
57 SPSS-X, SPSS Inc., Chicago, IL, USA.
58 Wishert, D., University College London, London, 1975, U.K.
59 Sepanova, Enskede, 1987, Sweden.
60 Wold, S., Technometrics, 20 (1978) 397–402.
61 Dunn III, W.J. and Wold, S., J. Med. Chem., 21 (1978) 922–930.
62 Wold, S., Sjöström, M., Carlson, R., Lundstedt, T., Hellberg, S., Skagerberg, B., Wikström, C. and Öhman, J., Anal. Chim. Acta, 191 (1986) 17–32.
63 Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S., J. Med. Chem., 30 (1987) 1126–1135.