A Novel Index for the Description of Molecular Linearity

Hiren Patel and Mark T. D. Cronin*

School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England

Received March 16, 2001

To investigate the description of molecular shape, a heterogeneous data set of 200 compounds was designed. The compounds included many disparate molecular shapes, and a total of 64 descriptors was calculated including topological indices, surface areas, molecular volumes, measures of molecular dimensions, and ratios of moments of inertia. Cluster analysis on the variables indicated that there are four major categories of steric descriptors, those for molecular bulk, dimensions, and two clusters representing varying aspects of shape. To describe molecular shape more efficiently, a novel linearity index based upon ratios of moments of inertia and the molecular weight is proposed.

INTRODUCTION

There are three main types of descriptors for use in the development of quantitative structure−activity relationships (QSARs): hydrophobic, electronic, and steric descriptors.[1,2] Of these, steric descriptors are associated with the size and shape of molecules. Molecular size and shape is important for many biological processes. This includes the ability of drug to bind at a receptor site. It is an important factor in the reactivity, and metabolism, of molecules, as molecule features associated with reactivity and metabolism may be sterically hindered. It may also be an important factor in the ability of a chemical to pass through a membrane. Long thin molecules may be more able to permeate through the gaps between cells than more spherical molecules. Molecular shape is also an important determinant in describing and quantifying chirality.[3]

The size and bulk of a molecule have been modeled successfully for use in QSAR analyses. They can be quantified easily using a number of descriptors, the most commonly used include molecular weight and molecular volume.[4] Molecular shape on the other hand, which refers to the distribution of molecular bulk according to the conformation, has been much less easy to quantify.[5] This has been especially the case for QSAR analyses, which often require a quantitative approach to assessing shape. There are a number of reasons for the difficulty in describing molecular shape. From a practical viewpoint, there are a number of three-dimensional conformations that a molecule may adopt. More fundamentally however, it is very difficult to assign a quantitative number to any particular shape. Further, there are many aspects to describing a molecular shape, e.g. by comparison to defined shapes such as ovals, spheres, etc., or more subjective estimates of the form of an object, such as the linear, or planar, nature of the molecule.

Thus one issue that has become apparent in the use of steric and shape descriptors in the development of QSARs is that while it is simple to describe molecular bulk, it is more difficult to ascribe a value for molecular shape. For instance, it is difficult to describe the difference between a linear molecule as compared to a nonlinear molecule. A good example is illustrated in Figure 1: hexane is clearly a linear molecule, cyclohexane is nonlinear. However both have very similar molecular weights, volumes, and surface areas.

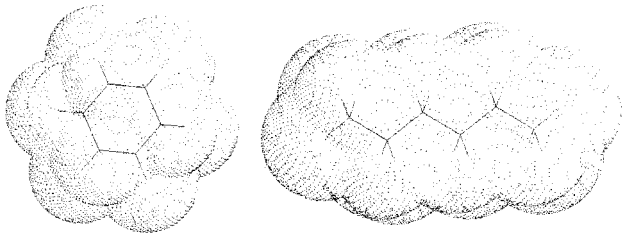

Figure 1. Three-dimensional representation of molecular shape to illustrate a linear molecule (hexane) as compared to a nonlinear molecule (cyclohexane).

To quantify molecular shape, a number of descriptors have been proposed. These include topological approaches such as the molecular connectivity[6] and kappa indices[7−9] and approaches based upon molecular dimensions including molecular shape analysis (MSA)[10] and STERIMOL parameters.[11] Quantitative methods to assess the shape of a molecule, which are based on molecular dimensions, do so by considering a combination of measurements. They usually take into account three point calculations on a molecule to determine a single parameter for molecular shape. While these are based on measures of the dimensions in particular molecular vectors, many do not describe shape well, particularly in terms of molecular linearity. Another approach has been the use of the Weighted Holistic Invariant Molecular (WHIM) descriptors.[12,13] These are based on a consideration of the x, y, and z coordinates of a molecule and scaled with differing weighting schemes.[12,13] Despite the applicability of these indices, it has not yet been demonstrated that they are able to describe the difference between linear and nonlinear molecules as shown in Figure 1.

Hypothetically, the linearity of molecule may be quantitated from a consideration of the dimensions of a molecule. There are a variety of approaches to calculate the dimensions

* Corresponding author phone: + 44 (0) 151 231 2066; fax: + 44 (0) 151 231 2170; e-mail: m.t.cronin@livjm.ac.uk.

10.1021/ci0103673 CCC: $20.00
Published on Web 08/01/2001

of a molecule; these attempt to find and assess the longest axis and axes perpendicular to it in 3-D space. A linear molecule therefore will have a longest axis considerably greater than axes perpendicular to it. A spherical molecule will have equal axes in these directions. Thus, it should follow that a ratio of these axes should provide a measure of how "linear" a molecule is. A caveat that must be borne in mind to the derivation of such an index for molecular linearity is that a large linear molecule will have a much greater ratio as compared to a small linear molecule. Thus an index for linearity may also need to be factored by a knowledge of relative molecular size.

The aim of this study, therefore, was to investigate descriptors for molecular shape, by which we took shape to mean the external form of the molecule. The study aimed to describe the linearity of a molecule in particular. Specifically, a theoretical data set of 200 compounds was devised to encompass a broad range of molecular shapes, from linear to nonlinear, highly branched, molecules. A variety of steric descriptors, which were easily available from commercial software, were calculated for these molecules and their inter-relationships assessed. Finally, a novel descriptor for molecular linearity was sought utilizing measures of molecular dimensionality.

METHODS

Selection of Chemical Structures. Two hundred chemical structures were selected for analysis. The criteria applied to select the chemicals aimed to create a data set with as much diversity in molecular shape as possible, with a particular emphasis to obtain a range of aromatic and aliphatic molecules that ranged from linear to nonlinear, branched (i.e. spherical) molecules. Chemicals were therefore included on a theoretical basis, rather than a selection of compounds from an existing database. The advantage of this approach is that structures may be selected to be representative solely of a variety of molecular shapes and so would not be biased by other criteria for data set design. Thus a wide range of structures was selected, including short to long chain unbranched alkanes; small to large cycloalkane structures; benzenes; naphthalenes; and branched and substituted derivatives of these. These provided considerable structural and shape diversity. Some structures from noncongeneric series (e.g. steroids) were also added to the data set in order to provide further diversity. A full list of the compounds selected is provided in Table 1.

Calculation of Physicochemical Properties. A total of 64 steric descriptors was calculated for each molecule. Initially, SMILES (Simple Molecular Input Linear Entry System) codes[14] were constructed for the each molecule in the data set and are reported in Table 1 (SMILES strings are available from the corresponding author). The SMILES strings were used as the input into the TSAR ver 3.2 molecular spreadsheet (Oxford Molecular Ltd.). These 2-D structures were converted into 3-D structures using the CORINA conformation analysis software in the TSAR molecular spreadsheet. Charges were assigned using the Charge-2 program, and the structure of compounds was roughly minimized using the COSMIC molecular mechanics force field. Subsequently, full energy minimization was performed using the AM1 Hamiltonian in the VAMP software (Oxford Molecular Ltd). Miscellaneous topological and steric descriptors were calculated from the energy minimized structures in TSAR. The individual indices calculated are listed in Table 2. The 3-D structures of the compounds were then saved as protein database (.pdb) files. These were converted into HyperChem (.hin) input files via the BABEL shareware file conversion system. These HyperChem files were used as the input to the Dragon software (free to download from the *Milano Research Group of Chemometrics* homepage at www.disat.unimib.it/chm/). WHIM shape descriptors were calculated using the DRAGON software as noted in Table 2. Following the calculation of the original descriptors, a further 16 descriptors were calculated as the ratios of the various moments of inertia. The moments of inertia for the length of the molecules were found to be more applicable and used to a greater extent in this study. These were included as it was hypothesized that the ratios of molecular dimensions would encode information regarding shape.

Statistical Analysis. All statistical analysis and data exploration was carried out using the MINITAB (ver 13.1) statistical software (MINITAB Inc.). Cluster analysis on variables was applied to evaluate the inter-relationships between the descriptors.

RESULTS AND DISCUSSION

A data set of 200 compounds was selected (see Table 1) for the analysis of descriptors for molecular shape and for attempted derivation of a novel index for molecular linearity. These compounds encompassed a range of molecular shape from linear alkanes, nonlinear branched alkanes, through to highly nonlinear branched ring structures and steroids. A total of 64 QSAR parameters considered to describe the steric attributes of the molecules were calculated for each compound.

Cluster analysis was performed on the 64 calculated variables and the novel linearity index (L_i) described below to investigate the inter-relationships and similarity between them. The results of the cluster analysis are shown as a dendrogram (see Figure 2). The interpretation of a cluster analysis is a subjective process; in this study there appear to be five major clusters or groups of clusters. These clusters are identified on the dendrogram and are labeled groups A−E in Figure 2. This indicates that the steric descriptors assessed can be classified into these five major types.

The significant cluster in Figure 2 (group A) is comprised mainly of molecular bulk descriptors. This is a tightly defined cluster, with considerable similarity between the descriptors. Predictably, molecular weight, molecular surface area, and molecular volume are all closely related. The zero- and first-order molecular connectivities are also related to molecular bulk. This has been noted previously by Dearden et al.[15] Interestingly, first-order kappa (K1) and kappa alpha (K1a) shape descriptors were also found to be descriptors of molecular bulk, despite the fact that they are both considered to be representations of the degree of complexity of the bonding arrangement of a molecule.[8] Kappa and kappa alpha differ only in the sense that kappa expresses this value on the assumption that all the atoms in the molecule are equivalent, whereas the kappa alpha terms are modified to incorporate the contribution of each atom to the overall shape

Table 1. SMILES String, Molecular Weight, K_u, and L_i of Compounds Considered Ordered According to L_i Value

ID no.	SMILES string[a]	MW	K_u	L_i
	Small Linear Molecules			
1	CF	34.04	0.245	6.842
2	CO	32.05	0.415	6.612
3	CCl	50.48	0.360	6.321
4	CC	30.08	0.372	6.164
5	C	16.05	0.000	5.909
6	O=CO	46.03	0.612	4.549
	Linear Molecules			
7	CCCO	60.11	0.559	3.585
8	CCCC	58.14	0.581	3.545
9	CCF	48.06	0.343	3.501
10	CCO	46.08	0.484	3.489
11	CCC	44.11	0.415	3.468
12	CCCCl	78.55	0.584	3.395
13	CCCCO	74.14	0.662	3.368
14	CCCCCO	88.17	0.724	3.364
15	CCCCCCO	102.20	0.781	3.325
16	CCCCC	72.17	0.664	3.308
17	CCCCCCCCO	130.26	0.851	3.304
18	CCCCCCCO	116.23	0.819	3.302
19	CCCCCC	86.20	0.742	3.300
20	CCCCCCCCCCO	158.32	0.893	3.293
21	CCCCCCCCCO	144.29	0.874	3.274
22	CCCCCCC	100.23	0.790	3.264
23	CCCCCCCCC	128.29	0.858	3.246
24	CCCCCCCCCCC	156.35	0.899	3.239
25	CCCCCCCC	114.26	0.831	3.238
26	CCCCCCCCCCCCC	184.41	0.924	3.236
27	CCCCCCCCCC	142.32	0.882	3.216
28	CCCCCCCCCCCC	170.38	0.913	3.206
29	CCCCCCl	106.61	0.741	3.204
30	CCCCCCCCCCCCCC	198.44	0.934	3.201
31	CCCCCCCCl	134.67	0.828	3.163
32	CCCCCCCCCl	148.69	0.856	3.119
33	CCCCCCCl	120.64	0.789	3.098
34	CCCCCl	92.58	0.670	3.066
35	CCCl	64.52	0.450	3.032
36	CI	141.94	0.350	2.979
	Linear Molecules with Branching on the Terminal Atom			
37	CCCCCCCCCC(=O)O	172.3	0.891	2.536
38	CCCCCCCCC(=O)O	158.27	0.871	2.472
39	CCCCCCCC(=O)O	144.24	0.848	2.440
40	O=N(=O)CCCCCCCCCCN(=O)=O	232.32	0.907	2.411
41	O=N(=O)CCCCCCCCCN(=O)=O	218.29	0.891	2.355
42	CCCCCCC(=O)O	130.21	0.815	2.346
43	O=N(=O)CCCCCCCCN(=O)=O	204.26	0.872	2.314
44	CCCCCC(=O)O	116.18	0.780	2.308
45	O=N(=O)CCCCCCCN(=O)=O	190.23	0.847	2.224
46	O=N(=O)CCCCCCN(=O)=O	176.2	0.816	2.188
47	CCCCC(=O)O	102.15	0.725	2.161
48	CCCC(=O)O	88.12	0.674	2.119
49	C1CC1	42.09	0.123	2.081
50	O=N(=O)CCCCCN(=O)=O	162.17	0.774	2.057
51	C(C)(C)CCCCCCCCC	170.38	0.881	2.043
52	O=N(=O)CCCCN(=O)=O	148.14	0.722	2.015
53	C(C)C(C)CCCCCCCC	170.38	0.876	1.851
54	CCC(=O)O	74.09	0.584	1.845
55	O=N(=O)CCCN(=O)=O	134.11	0.647	1.823
56	O=N(=O)CCN(=O)=O	120.08	0.564	1.763
57	CC(=O)O	60.06	0.568	1.674
58	CCI	155.97	0.338	1.609
59	C1CCC1	56.12	0.182	1.530
	Branched, Nonlinear, Molecules			
60	CCCCCc1ccccc1	148.27	0.762	1.395
61	C(C)C(C)C(C)CCCCCCC	184.41	0.842	1.370
62	O=N(=O)CN(=O)=O	106.05	0.438	1.367
63	Cc1ccc(C)cc1	106.18	0.629	1.358
64	C1CCC1C2CCC2	110.22	0.543	1.331
65	CCCc1ccccc1	120.21	0.651	1.317
66	CCCCc1ccccc1	134.24	0.690	1.294
67	Cc1ccccc1	92.15	0.480	1.261
68	CC(C)C(C)(C)CCCCCCCC	198.44	0.833	1.215

Table 1. (Continued).

ID no.	SMILES string[a]	MW	K_u	L_i
	Branched, Nonlinear, Molecules (Continued)			
69	CCc1ccccc1	106.18	0.527	1.201
70	CCCc1ccc(CCC)cc1	162.3	0.778	1.183
71	OCc1ccc(CO)cc1	138.18	0.686	1.182
72	C1CCCC1	70.15	0.253	1.168
73	CCc1ccc(CC)cc1	134.24	0.664	1.150
74	c1ccccc1	78.12	0.500	1.146
75	OC(=O)c1ccc(C(=O)O)cc1	166.14	0.696	1.115
76	C1CCCC1CC2CCCC2	152.31	0.644	1.104
77	COC(=O)c1ccc(C(=O)OC)cc1	194.20	0.840	1.087
78	CCC(CC)CCCCCCCCC	198.44	0.843	1.068
79	C1CCCCC1CCC2CCCCC2	194.40	0.731	0.968
80	CC(C)c1ccc(C(C)C)cc1	162.30	0.671	0.945
81	C1CCCCC1	84.18	0.267	0.937
82	c1cccc2ccccc12	128.18	0.500	0.917
83	c1ccccc1CCCc2ccccc2	196.31	0.697	0.909
84	CC(C)C(C)(C)C(C)(C)CCCCCCC	226.50	0.786	0.878
85	Cc1c(C)cccc1	106.18	0.455	0.878
86	CCCCc1ccc(CCCC)cc1	190.36	0.767	0.878
87	C1C(C)CC1c2ccc(C3CC(C)C3)cc2	214.38	0.802	0.874
88	CCCCc1ccc(CCCCC)cc1	218.42	0.816	0.836
89	C1CCCCCC1	98.21	0.257	0.834
90	CCCC(CC)CCCCCCCC	198.44	0.743	0.725
91	C1CCCCCC1CCCC2CCCCCC2	236.49	0.711	0.698
92	Cc1c(C)c(C)ccc1	120.21	0.449	0.695
93	Cc1c(C)c(C)c(C)cc1	134.24	0.448	0.689
94	CC(C)C(C)(C)C(C)(C)C(C)(C)CCCCCC	254.56	0.740	0.688
95	Cc1c(C)ccc2ccccc12	156.24	0.487	0.676
96	CCC(C)C(C)C(C)C(C)C(C)CCCC	226.50	0.677	0.672
97	C1CCCCCCC1	112.24	0.240	0.670
98	C1CCCCCCC1CCCCC2CCCCCCC2	278.58	0.800	0.661
99	CCC(C)C(C)C(C)CCCCCC	198.44	0.684	0.650
100	Cc1cccc2ccccc12	142.21	0.485	0.645
101	CCCC(CCC)CCCCCCCCC	226.50	0.773	0.623
102	C1CCCCCCCCC1	140.30	0.325	0.600
103	CCC(C)C(C)C(C)C(C)C(C)C(C)CCC	240.53	0.630	0.580
104	C1CCCCCCCC1	126.27	0.288	0.571
105	Cc1c(C)c(C)cc2ccccc12	170.27	0.472	0.570
106	Cc1c(C)c(C)c(C)c(C)c1	148.27	0.448	0.554
107	CCC(C)C(C)C(C)C(C)CCCCC	212.47	0.697	0.550
108	CCc1c(CC)cccc1	134.24	0.355	0.549
109	CCC(CC)C(CC)(CC)CCCCCCCC	254.56	0.743	0.532
110	CCc1cccc2ccccc12	156.24	0.437	0.522
111	CCCCc1cccc2ccccc12	184.30	0.550	0.510
112	CCCc1cccc2ccccc12	170.27	0.484	0.505
	Highly Branched, Nonlinear, Molecules			
113	CCC(C)C(C)C(C)C(C)C(C)C(C)C(C)CC	254.56	0.629	0.485
114	CCc1c(CC)ccc2ccccc12	184.30	0.428	0.473
115	CCCCC(CCC)CCCCCCCC	226.50	0.677	0.461
116	Cc1c(C)c(C)c(C)c2ccccc12	184.30	0.468	0.452
117	CCc1c(CC)c(CC)ccc1	162.30	0.312	0.449
118	CC(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)CCCCC	282.62	0.645	0.442
119	Cc1c(C)c(C)c(C)c(C)c1(C)	162.30	0.447	0.441
120	CCC(C)C(C)C(C)C(C)C(C)C(C)C(C)C(C)C	268.59	0.631	0.438
121	Cc1c(C)c(C)c(C)c2c(C)c(C)ccc12	212.36	0.467	0.420
122	CCCC(CC)C(CC)CCCCCCC	226.50	0.585	0.411
123	Cc1c(C)c(C)c(C)c2c(C)cccc12	198.33	0.466	0.405
124	CC(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)CCCC	310.68	0.630	0.405
125	CCc1c(CC)c(CC)c(CC)cc1	190.36	0.378	0.404
126	CC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CCC34C	330.51	0.577	0.399
127	Cc1c(C)c(C)c(C)c2c(C)c(C)c(C)cc12	226.39	0.467	0.392
128	CCCc1c(CCC)cccc1	162.30	0.395	0.387
129	CCCCC(CCCC)CCCCCCCCC	254.56	0.658	0.370
130	COC(=O)CCCCCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	518.71	0.790	0.364
131	CC(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)CCC	338.74	0.601	0.359
132	CCCc1c(CCC)ccc2ccccc12	212.36	0.411	0.355
133	CCCCCCCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	488.73	0.790	0.353
134	CCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	418.58	0.633	0.351
135	CCCCCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	460.67	0.737	0.349
136	CCc1c(CC)c(CC)cc2ccccc12	212.36	0.364	0.343
137	CC(C)C(C)C(C)C(C)C(C)C(C)C(C)C(C)C(C)C(C)C(C)CC	366.80	0.596	0.342
138	Cc1c(C)c(C)c(C)c2c(C)c(C)c(C)c(C)c12	240.42	0.409	0.338
139	CCCc1c(CCC)c(CCC)ccc1	204.39	0.379	0.336

Table 1. (Continued)

ID no.	SMILES string[a]	MW	K_u	L_i
	Highly Branched, Nonlinear, Molecules (Continued)			
140	CCCCc1c(CCCC)cccc1	190.36	0.391	0.329
141	COC(=O)CCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	476.62	0.718	0.322
142	CC13CCC(=O)C=C1CCC4C2CCC(O)(C(=O)COC(=O)CCC(N)=O)C2(C)C(O)CC34	461.61	0.712	0.321
143	CN(C)C(=O)CCC(=O)OCC(=O)C3(O)CCC4C2CCC1=CC(=O)CCC1(C)C2CC(O)C34C	489.67	0.737	0.317
144	CC13CCC(=O)C=C1CCC4C2CCC(O)(C(=O)COC(=O)CCCCCO)C2(C)C(O)CC34	476.67	0.775	0.317
145	CC13CCC(=O)C=C1CCC4C2CCC(O)(C(=O)COC(=O)CCC(O)=O)C2(C)C(O)CC34	462.59	0.701	0.316
146	CC13CCC(=O)C=C1CCC4C2CCC(O)(C(=O)COC(=O)CCCCCC(N)=O)C2(C)C(O)CC34	503.70	0.801	0.315
147	CC13CCC(=O)C=C1CCC4C2CCC(O)(C(=O)COC(=O)CCCCCC(O)=O)C2(C)C(O)CC34	504.68	0.792	0.313
148	CCC(CC)C(CC)(CC)C(CC)(CC)CCCCCCC	310.68	0.595	0.313
149	CCCCCC(CCCC)CCCCCCCC	254.56	0.581	0.309
150	CCCC(CC)C(CC)C(CC)CCCCCC	254.56	0.502	0.303
151	CCc1c(CC)c(CC)c(CC)c(CC)c1	218.42	0.319	0.301
152	CCCC(CC)C(CC)C(CC)C(CC)C(CC)CCCC	310.68	0.513	0.292
153	CCCc1c(CCC)c(CCC)c(CCC)cc1	246.48	0.434	0.285
154	CCCC(CCC)C(CCC)(CCC)CCCCCCCC	310.68	0.603	0.283
155	CC(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C(C)(C)C	394.86	0.521	0.276
156	CCCC(CC)C(CC)C(CC)C(CC)CCCCC	282.62	0.557	0.271
157	CCCCc1c(CCCC)ccc2ccccc12	240.42	0.390	0.267
158	CCCC(CC)C(CC)C(CC)C(CC)C(CC)C(CC)CCC	338.74	0.459	0.254
159	CCc1c(CC)c(CC)c(CC)c2ccccc12	240.42	0.367	0.253
160	CNC14CCC25C(Oc3c(O)ccc(C1)c23)C(=O)CCC45	285.37	0.332	0.247
161	CCCCc1c(CCCC)c(CCCC)ccc1	246.48	0.332	0.246
162	CCC(CC)C(CC)(CC)C(CC)(CC)C(CC)(CC)CCCCCC	366.80	0.477	0.244
163	CCCCC(CCC)C(CCC)CCCCCCC	268.59	0.449	0.243
164	CCc1c(CC)c(CC)c(CC)c2c(CC)cccc12	268.48	0.370	0.242
165	CCc1c(CC)c(CC)c(CC)c2c(CC)c(CC)ccc12	296.54	0.384	0.240
166	CCCc1c(CCC)c(CCC)cc2ccccc12	254.45	0.364	0.238
167	CCc1c(CC)c(CC)c(CC)c(CC)c1(CC)	246.48	0.336	0.234
168	CCc1c(CC)c(CC)c(CC)c2c(CC)c(CC)c(CC)cc12	324.60	0.371	0.213
169	CCCC(CC)C(CC)C(CC)C(CC)C(CC)C(CC)C(CC)CC	366.80	0.344	0.203
170	CCCCc1c(CCCC)c(CCCC)c(CCCC)cc1	302.60	0.394	0.201
171	CCCc1c(CCC)c(CCC)c(CCC)c(CCC)c1	288.57	0.353	0.200
172	CCC(CC)C(CC)(CC)C(CC)(CC)C(CC)(CC)C(CC)(CC)CCCCC	422.92	0.444	0.195
173	CCCCC(CCC)C(CCC)C(CCC)CCCCCC	310.68	0.355	0.187
174	CCCc1c(CCC)c(CCC)c(CCC)c2ccccc12	296.54	0.382	0.185
175	CCc1c(CC)c(CC)c(CC)c2c(CC)c(CC)c(CC)c(CC)c12	352.66	0.373	0.181
176	CCCCc1c(CCCC)c(CCCC)cc2ccccc12	296.54	0.265	0.179
177	CCCC(CCC)C(CCC)(CCC)C(CCC)(CCC)CCCCCCC	394.86	0.394	0.175
178	CCCc1c(CCC)c(CCC)c(CCC)c2c(CCC)cccc12	338.63	0.387	0.173
179	CCCCC(CCCC)C(CCCC)(CCCC)CCCCCCCC	366.80	0.461	0.172
180	CCCCC(CCC)C(CCC)C(CCC)C(CCC)C(CCC)CCCC	394.86	0.393	0.165
181	CCCc1c(CCC)c(CCC)c(CCC)c2c(CCC)c(CCC)ccc12	380.72	0.399	0.163
182	c1ccccc1C(c2ccccc2)CC(c4ccccc4)c3ccccc3	348.51	0.237	0.161
183	CCCCC(CCC)C(CCC)C(CCC)C(CCC)CCCCC	352.77	0.365	0.159
184	CCCCCC(CCCC)C(CCCC)CCCCCCC	310.68	0.281	0.157
185	CCCc1c(CCC)c(CCC)c(CCC)c(CCC)c1(CCC)	330.66	0.374	0.152
186	CCCCc1c(CCCC)c(CCCC)c(CCCC)c2ccccc12	352.66	0.312	0.150
187	CCCCc1c(CCCC)c(CCCC)c(CCCC)c(CCCC)c1	358.72	0.280	0.139
188	CCCc1c(CCC)c(CCC)c(CCC)c2c(CCC)c(CCC)c(CCC)cc12	422.81	0.389	0.138
189	CCCCc1c(CCCC)c(CCCC)c(CCCC)c2c(CCCC)cccc12	408.78	0.372	0.131
190	CCCC(CCC)C(CCC)(CCC)C(CCC)(CCC)C(CCC)(CCC)CCCCCC	479.04	0.301	0.127
191	CCCCCC(CCCC)C(CCCC)C(CCCC)CCCCCC	366.80	0.301	0.121
192	CCCCC(CCCC)C(CCCC)(CCCC)C(CCCC)(CCCC)CCCCCCC	479.04	0.354	0.120
193	CCCCc1c(CCCC)c(CCCC)c(CCCC)c2c(CCCC)c(CCCC)ccc12	464.90	0.354	0.118
194	CCCCC(CCC)C(CCC)C(CCC)C(CCC)C(CCC)C(CCC)CCC	436.95	0.298	0.113
195	CCCc1c(CCC)c(CCC)c(CCC)c2c(CCC)c(CCC)c(CCC)c(CCC)c12	464.90	0.393	0.113
196	CCCCc1c(CCCC)c(CCCC)c(CCCC)c(CCCC)c1(CCCC)	414.84	0.316	0.108
197	CCCCCC(CCCC)C(CCCC)C(CCCC)C(CCCC)CCCCC	422.92	0.307	0.106
198	c1ccccc1CC(Cc2ccccc2)(Cc3ccccc3)Cc1ccccc1	376.57	0.050	0.106
199	CCCCCC(CCCC)C(CCCC)C(CCCC)C(CCCC)C(CCCC)CCCC	479.04	0.260	0.102
200	c1ccccc1C(c2ccccc2)(c3ccccc3)CC(c4ccccc4)(c5ccccc5)c6ccccc6	500.71	0.093	0.095

[a] SMILES strings are available electronically from the corresponding author.

of an atom, as compared to a carbon sp^3 atom.[9] Despite this distinction kappa and kappa alpha are themselves highly related. The final set of descriptors in this cluster was the calculated molecular polarizabililty and its components in the X, Y, and Z directions (PXX, PYY, and PZZ). Molecular polarizability is normally considered to be a function of molecular size,[1] so there is no surprise in their appearing in this cluster. The predominance of a large cluster of descriptors for molecular size has also been shown by Basak et al.[16] in a cluster analysis of topological indices.

Closely related to the descriptors of bulk are a number of small clusters which have been described together as group B. This group appears to represent descriptors of varying aspects of molecular dimensions. The clearest indication of

Table 2. Names, Abbreviations, and Source of Molecular Descriptors Considered in This Study

descriptor(s)	abbreviation(s)	source
mean polarizibility and its components	MP, PXX, PYY, PZZ, PXY, PXZ, PYZ	VAMP (TSAR)
total energy	TE	VAMP (TSAR)
surface area	SA	VAMP (TSAR)
molecular mass	MW	TSAR
molecular volume	MV	TSAR
ellipsoidal volume	EV	TSAR
kappa shape indices	K1, K2, K3	TSAR
kappa alpha shape indices	Ka1, Ka2, Ka3	TSAR
shape flexibility	SF	TSAR
molecular flexibility	Phi	TSAR
Kier connectivity indices (atoms, bonds) zero and first order	Chi0, ChiOv, Chi1, Chi1v	TSAR
inertia moments (length)	IM1(L), IM2(L), IM3(L),	TSAR
inertia moments (size)	IM1(S), IM2(S), IM3(S)	TSAR
ratios of inertia moments (length)	IM1/2(L), IM1/3(L), IM2/3(L), IM2/1(L), IM3/1(L), IM3/2(L), IM1/2/3(L), IM3/2/1(L), IM2/1/3(L), IM1 × 2 × 3(L)	TSAR
ratios of inertia moments (size)	IM1/2(S), IM1/3(S), IM2/3(S), IM2/1(S), IM3/1(S), IM3/2(S)	TSAR
WHIM 1st component direction shape index	$P1_{u,m,v,e,p,s}$[a]	Dragon
WHIM 2nd component direction shape index	$P2_{u,m,v,e,p,s}$[a]	Dragon
WHIM global shape index	$K_{u,m,v,e,p,s}$[a]	Dragon

[a] K_x; $P1_x$; $P2_x$; where x is the weighting scheme.

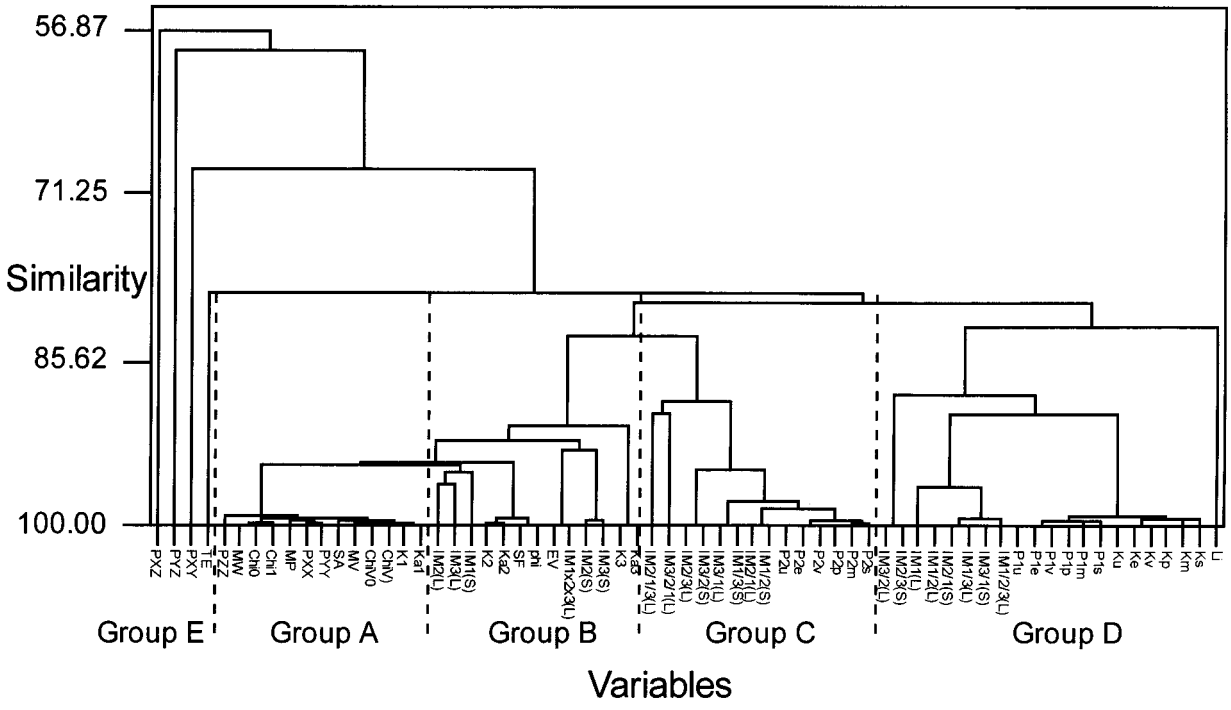

Figure 2. Dendrogram from cluster analysis of steric desccriptors considered in this study.

this is the presence of the raw moments of inertia (IM1(S); IM2(S); IM3(S); IM2(L); IM3(L)). Other kappa indices (K2; Ka2; K3; Ka3) not only are more loosely clustered to these raw dimensions indicating some similarity but also that they may encode other information as well. This is not surprising, as these are topological indices and so will not contain exactly the same information as dimensions derived from the 3-D molecular structure. K2 and Ka2 give a measure of the degree of linearity or starlikeness of bonding patterns, whereas K3 and Ka3 indicate the degree of branching at the center of a molecule.[8,9] Thus it may be considered that the second- and third-order kappa indices are included in group B (as opposed to the first-order kappa indices in group A), since they represent information regarding more specific molecular dimensions. The descriptors for shape flexibility (SF), molecular flexibility (phi), and ellipsoidal volume (EV) are also present within group B, again confirming their relationship to molecular dimensionality.

The WHIM shape descriptors (K_x; $P1_x$; $P2_x$; where x is the weighting scheme) and the ratios of the moments of inertia are distributed between groups C and D in Figure 2. Taking each group in turn, group D is effectively made up to two clusters and one single variable. The first cluster is tight and contains the WHIM descriptors K_x and $P1_x$. A looser

cluster in this group contains various ratios of the moments of inertia. Also loosely associated with this group is the linearity index explained below. Closer examination of the variables in this group suggests that they are representative of molecular linearity (the meaning and interpretation of this group is discussed below). Group C contains a tight cluster of WHIM descriptors ($P2_x$) and more ratios of moments of inertia more loosely clustered. Thus, group C is considered to describe molecular dimensions. Due to the similarity of groups C and D, the discussion of them will be considered together.

Both groups C and D contain a number of WHIM descriptors. These are calculated from a principal component analysis of the molecular coordinates of the 3-D structure of a compound.[12,13] There are three WHIM descriptors of shape (K_x; $P1_x$; $P2_x$). P1 and P2 (called directional WHIM descriptors) are first and second component directional shape indices, respectively. They are derived from eigenvalue proportions that are related to molecular size. The K series, derived from the directional WHIM descriptors, provide an overall global view of the shape of a molecule. The *K* value for the ideal straight molecule should be such that $K = 1$, for a perfectly spherical molecule $K = 0$, and for all planar molecules $K = 0.5-1$. All WHIM descriptors may be unweighted (U) or weighted using schemes for atomic masses (M); the van der Waals volumes (V); the Mulliken atomic electronegativites (E); the atomic polarizabilities (P); and the electrotopological indices of Kier and Hall (S).[12,13] The weighting scheme provides a means of searching for the principal axes with respect to atomic property. For the purposes of this study only the "unweighted" descriptors were used for further analyses. The weighting schemes mean that the tight clustering of the six variants of K_x, $P1_x$, and $P2_x$ in groups C and D is expected. The advantage of using WHIM descriptors over topological descriptors is that they are able to distinguish between different conformations and geometric isomers of the same molecule.[12,13] This thus eliminates the need to align the molecules before calculation of any parameters, as is the case with many 3-D molecular descriptors (such as those from CoMFA[17]). In both groups C and D, the ratios of the moments of inertia are strongly related to the WHIM shape descriptors. This also suggests that these ratios are strongly associated with aspects of molecular shape.

Group E is the loosest cluster that contains only four miscellaneous variables. They do not share any significant similarity to each other or any apparent similarity to any of the other descriptors. The variables are three of the polarizability components (PXY, PXZ, PYZ) and total molecular energy (TE). While the polarizability components are based on a fundamental descriptor of molecular size, these variants appear to have little relation to this property.

To investigate the description of shape, and in particular molecular linearity, in more detail, the WHIM descriptors were examined. As mentioned above K is considered to be an index of linearity, and the unweighted *K* value (K_u) was taken as the most fundamental descriptor of linearity. To explore the relative meaning of K_u, Figure 3 shows the K_u values for the compounds in the data set plotted against molecular weight. The K_u values for the compounds in the data set are also noted in Table 1.

Figure 3 indicates that there is no collinearity between K_u and molecular size. Highlighted (with a closed circle) in

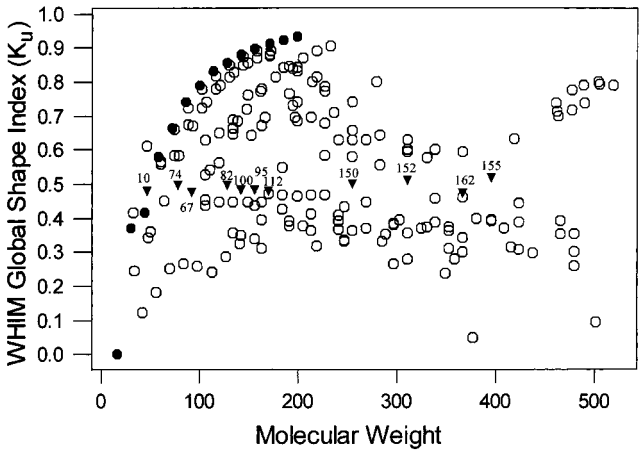

Figure 3. Plot of WHIM global shape descriptors (K_u) against molecular weights (MW) of compounds in the data set. Legend: solid circles are the congeneric series of alkanes. Solid triangles are a selection of compounds with K_u of approximately 0.5 (refer to text for further information). Numbers relate to the ID number of the chemicals in Table 1.

Figure 3 are the K_u values for a set of 14 straight chain alkanes, starting with methane. These were chosen as they represent a series of molecules which are linear but increase in size. As one would expect, methane has a zero value as its shape can be described as being spherical, rather than linear. As the number of carbons increases there is also an increase in K_u as the alkane chains become "more linear". Thus the transition from the spherical to the linear molecule category is described by K_u. The value of K_u for the largest linear alkane considered in this analysis (comprising of 14 carbon atoms) is very close to one, implying a high degree of molecular linearity. However, in this case molecular linearity and size may be collinear. Further examination of Figure 3 indicates this deficiency associated with the global shape descriptor K. While K is able to differentiate successfully between spherical and highly linear molecules, it is less capable of differentiating between other molecular shapes. Eleven molecules are highlighted in Figure 3 (with a solid triangle) that have a K_u value very close to 0.5. They have a range of different shapes and sizes, ranging from ethanol to long chain branched alkanes, and benzene to branched naphthalene derivatives. Thus the K_u descriptor is not able to differentiate these varied shapes. In addition, the K_u values are scaled between 0 and 1, thus their usefulness in QSAR analysis to describe shape may be restricted.

To improve upon the K_u descriptor for molecular shape, and in particular, linearity, a consideration of molecular dimensions was made. The moments of inertia have been calculated in two forms, as moments of inertia based on size and on length. Calculation of these descriptors for a molecule is accomplished using the so-called "inertia tensor".[18] The moments of inertia describe the size of a molecule and are calculated as the mass distribution in three directions: x, y, and z. The principal axes of inertia are calculated as "three mutually perpendicular directions" each associated with a principal moment of inertia with an orthogonal conformation. Figure 2 shows that IM2 and IM3 are very closely related; however, IM1 is independent of these values. It also indicates that some of the ratios calculated from the moments of inertia for lengths (in group D) are related to WHIM descriptors thought to describe molecular shape. Based upon this

Table 3. Main Groups of Compounds Based on L_i as Shown in Figure 4

L_i range	description	examples
4−7	small linear molecules	methane, ethane, methanol, chloromethane
3−4	straight chain linear molecules	linear alkanes, alcohols, haloalkanes
1.5−3	straight chain linear molecules with terminal branching	carboxylic acids, nitroalkanes
0.5−1.5	branched nonlinear molecules	aromatics, multiple branched alkanes, cycloalkanes
0−0.5	highly branched nonlinear molecules	steroids, aromatics, multiple branched alkanes, cycloalkanes

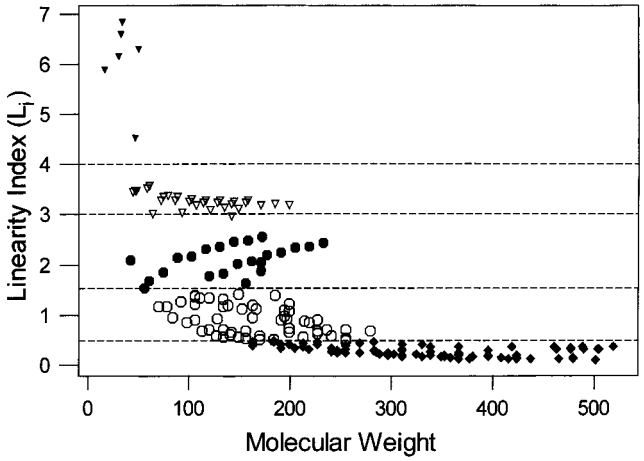

Figure 4. Plot of linearity index (L_i) against molecular weights (MW) of compounds in the data set, showing distribution of compounds according to L_i as noted in Table 3.

knowledge, and the hypothesis that molecular linearity could be quantified using ratios of molecular dimensions, a novel index is proposed. This is calculated using a similar methodology as used for the calculation of the global K_u shape descriptor. In order that the relative size of a molecule is represented in the index, the calculation also incorporates a squared term for molecular weight to emphasize the contribution of the largest molecules. Trial and error showed that the square root of the index provided a better discrimination of shape. The whole index was multiplied by a factor of 100 to obtain a more manageable scale. The formula used to calculate the "linearity index" (L_i) is as follows:

$$L_i = \sqrt{\frac{IM1(L)/IM2(L)/IM3(L)}{MW^2}} \times 100 \qquad (1)$$

To investigate the meaning of L_i, the values were plotted against the molecular weight of the compounds (see Figure 4). As with the WHIM shape descriptor, L_i is not related to molecular size. Closer examination of Figure 4 allows it to be divided into five sections. Each section, each representing a range of L_i values, relates to different levels of molecular linearity. The sections are summarized in Table 3 and are described in more detail below.

The first group of compounds in Figure 4 have a L_i value between 4 and 7. These compounds are small (comprised of one, two, or three heteroatoms) molecules. The second group of compounds have a range of L_i values from 3 to 4. These compounds are highly linear i.e., unbranched alkanes. The molecules in this group include not only straight chain alkanes but also all other linear compounds, including aliphatic alcohols and haloalkanes. Indeed, all linear molecules in the data set were present in this group. This is favorable as compared to the WHIM K_u descriptor, which was allocated a range of values for these linear compounds (see Figure 3) which are in common with a whole range of other molecules.

The third group of compounds have a range of L_i values between 1.5 and 3. These compounds consist predominantly, with the exception of two molecules, of linear molecules with branching at one or both of its terminal carbon(s). Thus, included are compounds such as aliphatic carboxylic acids and nitroalkanes. The other two compounds found within this range were cyclopropane and cyclobutane. Since none of the larger cycloalkanes are in this range, one can assume that this range of L_i values accounts only for small ring structures.

Compounds with L_i values between 0.5 and 1.5 were found in the fourth group. These compounds may be considered to be nonlinear and have a limited number of branches, they include a more heterogeneous selection of structures. The branches on the compounds are short in length (typically one or two heteroatoms) and are in all positions on the molecules, as opposed to only terminal branching on the molecules with L_i values between 1.5 and 3. Included are linear alkanes with multiple branching with small substituents or single or double branching with larger substituents. Simple aromatic compounds are also found in this group, with little or no branching.

The final group of compounds has a range of L_i values from 0 to 0.5. It contains compounds with structures that are highly nonlinear and branched. The compounds in this range extend from those in the previous group in terms of complexity of branching, including molecules consisting of multiple fused rings, e.g. steroids and highly branched naphthalenes.

L_i appears to provide a quantitative measure that separates linear molecules (such as the alkanes) from nonlinear and branched molecules (such as branched alkanes, aromatic compounds, and steroids). L_i provides this separation without being influenced by molecular size, providing a considerable advantage of the use of the WHIM K descriptors. Further, examination of Figure 2 confirms that the linearity index is not closely associated with other steric descriptors and is encoding different information about the molecular properties.

When attempting to quantify molecular shape, it is important to appreciate that it is not possible to describe it quantitatively in terms of one descriptor. The approach that must be taken is to identify elements of shape, such as bulk, dimensions, linearity, and use these appropriately in QSAR analyses. The descriptor proposed, L_i, allows for the quantitative description of one of these elements of molecular shape. It should be emphasized, therefore, that L_i will not quantify shape *in toto*. Other approaches to describe shape do so by comparison with other molecules (as opposed to L_i which is independent of other molecules). Obviously a comparison of molecular shape is essential for phenomena such as receptor binding. Techniques suitable to address

shape and conformational flexibility include Catalyst/SHAPE,[19] Compass,[20] and the 4D-QSAR approach.[21] Briefly, Catalyst/SHAPE allows for the generation of lead compounds in drug discovery by searching 3-D databases for shape similarity between molecules and identifies molecules with similar 3-D shapes.[19] Compass also identifies molecules with similar shapes, using a neural network to predict molecules with suitable conformations and alignments.[20] The 4D-QSAR approach considers both conformation and alignment freedom (the sample of which constitutes the "fourth" dimension) in the construction of 3-D QSAR models.[21] Such techniques are distinct from the aim of this study to quantify, independently of other molecules, a particular aspect of molecular shape, namely linearity.

In conclusion, a large number of steric descriptors have been calculated from commercially and freely available software packages. Many of these describe molecular size and bulk. Other descriptors are found to encode information relating to molecular dimensions, and shape. Of these, the WHIM K_u descriptor describes some aspects of linearity and size. To describe molecular linearity more explicitly, a novel descriptor is proposed, the linearity index (L_i). The linearity index appears to discriminate well linear molecules from nonlinear and branched molecules, regardless of molecular size. Further studies will investigate the application of the linearity index to predict endpoints such as skin permeability, where the shape of molecule may determine a molecule's ability to pass through gaps between cells.

ACKNOWLEDGMENT

This study was funded in part through the European Chemical Industry Council (CEFIC) Long-Range Initiative (LRI).

REFERENCES AND NOTES

(1) Dearden, J. C. Physico-Chemical Descriptors. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; EEC: Brussels, 1990; pp 25−59.

(2) Livingstone, D. J. The Characterization of Chemical Structures using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(3) Benigni, R.; Cotta-Ramusino, M.; Gallo, G.; Giorgi, F.; Giuliani, A.; Vari, M. R. Deriving a Quantitative Chirality Measure from Molecular Similarity Indices. *J. Med. Chem.* **2000**, *43*, 3699−3703.

(4) Cronin, M. T. D. Molecular Descriptors of QSAR. In *Proceedings of the Seminar of Current Topics in Toxicology: QSAR in Toxicology*; Coccini, T., Giannoni, L., Karcher, W., Manzo, L., Roi, R., Eds.; Commission of the European Communities: Luxembourg, 1992; pp 43−54.

(5) Tute, M. S. History and Objectives of Quantitative Drug Design. In *Comprehensive Medicinal Chemistry. Volume 4. Quantitative Drug Design*; Ramsden, C. A., Ed.; Pergammon Press: Oxford, 1990; pp 1−31.

(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: Chichester, England, 1986.

(7) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(8) Kier, L. B. Shape Indexes of Order One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

(9) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.

(10) Hopfinger, A. J. A QSAR Study of the Ames Mutagenicity of 1-(X-Phenyl)-3,3-dialkyltriazenes using Molecular Potential Energy Fields and Molecular Shape Analysis. *Quant. Struct.-Act. Relat.* **1984**, *3*, 1−5.

(11) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. 7, pp 165−207.

(12) Todeschini, R.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−119.

(13) Todeschini, R.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. 6. Application of WHIM Descriptors in QSAR Studies. *Quant. Struct.-Act. Relat.* **1997**, *16*, 120−125.

(14) Weininger, D. SMILES, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(15) Dearden, J. C.; Bradburne, S. J. A.; Cronin, M. T. D.; Solanki, P. The Physical Significance of Molecular Connectivity. In *QSAR 88 − Proceedings of the Third International Workshop on Quantitative Structure−Activity Relationships in Environmental Toxicology*; Turner, J. E., England, M. W., Schultz, T. W., Kwaak, N. J., Eds.; USDOE: Oak Ridge, TN, 1988; pp 43−50.

(16) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.

(17) Cramer, R. D. III.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(18) Riley, K. F. *Mathematical Methods for the Physical Sciences*; Cambridge University Press: Cambridge, 1973.

(19) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80−86.

(20) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: a Shape-Based Machine Learning Tool for Drug Design. *J. Comput.-Aided Mol. Design* **1994**, *8*, 635−652.

(21) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

CI0103673