Electronic van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure−Activity Correlations in Olfactory Databases

Barry K. Lavine* and Charles E. Davidson

Department of Chemistry, Clarkson University, Potsdam, New York 13699-5810

Curt Breneman and William Katt

Department of Chemistry, Rensselaer Polytechnic Institute, Troy, New York 12180

Received May 21, 2003

A methodology to facilitate the intelligent design of new odorants (e.g., musks) with specialized properties has been developed as part of an ongoing research effort in machine learning. In a traditional framework, the introduction of a new odorant is a lengthy, costly, and laborious discovery, development, and testing process. We propose to streamline this process utilizing large existing olfactory databases available through the open scientific literature as input for a new structure/activity correlation methodology. The first step in this process is to characterize each molecule in the database by an appropriate set of descriptors. To accomplish this task, an enhanced version of Breneman's Transferable Atom Equivalent (TAE) descriptor methodology will be used to create a large set of electron density derived shape/property hybrid (PEST), wavelet coefficient (WCD), and TAE histogram descriptors. We have chosen these molecular property descriptors to represent the problem because they have been shown to contain pertinent shape and electronic properties of the molecule and correlate with key modes of intermolecular interactions. Traditional QSAR methodologies, which employ fragment based descriptors, have been shown to be effective for QSAR development within homologous sets of molecules but are less effective when applied to data sets containing a great deal of structural variation. In contrast to previous attempts at SAR, our use of shape-aware electron density based molecular property descriptors has removed many of the limitations brought about by the use of descriptors based on substructure fragments, molecular surface properties, or other whole molecule descriptors. Another reason for the mixed success of past QSAR efforts can be traced to the nature of the underlying modeling problem, which is often quite complex. To meet these challenges, a genetic algorithm for pattern recognition analysis has been developed that selects descriptors which create class separation in a plot of the two largest principal components of the data while simultaneously searching for features that increase clustering of the data.

INTRODUCTION

The chemical sense of olfaction is a complex and poorly understood phenomenon. While it is an integral part of everyday life, information about the relationship between chemical structure and odor is scarce. For a compound to have an odor, it is generally agreed that it must be volatile as well as both lipid and water-soluble. Beyond this general description of characteristics, there is no agreement among researchers as to which molecular properties and structural features are responsible for the olfactory impressions invoked by odorants.

Several theories have been proposed that attempt to correlate molecular structure with the perceived odor quality of a substance. Wright[1] has suggested that odor quality could be predicted on the basis of low-energy molecular vibrations occurring below 600 cm^{-1}. Correlations between spectra and odor in that range have been previously reported, but they were often achieved by dismissing discrepant peaks as anosmic and adding some computed "difference" frequencies, particularly among musks and camphoraceous odorants.[2] Beets[3] discussed odor quality relative to molecular shape as represented by oriented profiles and functional groups. For example, most molecules that smell like camphor are spherical. However, the converse statement that all spherical molecules smell like camphor is not true. Evidently, there are other factors, which must be taken into account. Theimer and Davies[4] have discussed the importance of molecular cross-sectional areas and free energies of desorption in relation to odor. Amoore[5] has reported that molecular shape, size, and electronic nature of a molecule are correlated to odor quality. Dravnieks and Laffort[6] have related odor quality to intermolecular forces. Axel[7] has reported that perception and recognition of odor is mediated by G-protein coupled transmembrane protein receptors located on the surface of cilia cells at the olfactory epithelium.

Analysis of structure−activity relationships (SARs) using computer assisted methods and pattern recognition techniques can provide a practical approach to the study of odorants. The heart of this approach is finding a set of molecular descriptors from which discriminating relationships can be found. According to the current theories of olfaction, the perception of odor is initiated by the interaction of the odorant with the olfactory receptor sites in the nose. Olfactory excitation only occurs if the size and shape of the stimulant is the complement of the receptor or if the stimulant possesses sufficient conformational flexibility to attain the

10.1021/ci030016j CCC: $25.00 © 2003 American Chemical Society
Published on Web 10/09/2003

correct shape. The spatial arrangement of the stimulant's functional and steric groups must also conform to the overall three-dimensional geometry of the receptor. It is logical to apply this information to a structure-olfaction study during the key step: the development of molecular descriptors. However, only topological and bulk geometric descriptors—molecular connectivity indices, substructures, substructural molecular connectivity environment descriptors, molecular volume, and principal moments of inertia—have been used to describe molecular shape in previously published SAR studies of musks and other odorants.[8−15] Descriptors, which contain information about the olfactory process, need to be developed and tested in order to formulate more effective SARs.

We have been studying musk odorants and structurally related nonmusk compounds using computer generated molecular descriptors and pattern recognition techniques for several reasons. First, musks arouse considerable industrial interest because of their fixative properties. Almost all fragrances sold commercially contain musk. Second, a wealth of information on musks is available in the public literature. Third, musks are interesting from a structure/activity point of view since they contain a variety of different structural types thereby making musk odor prediction more challenging. Fourth, musks have a distinct odor, which is rarely confused with any other odor. Such reliability in the assignment of odor yields a well-characterized data set with few mislabeled compounds.

In a previously published study,[16] we utilized nonparameteric linear discriminants to differentiate musk from nonmusk odorants based on a set of computer generated molecular structure descriptors. A training set of 148 bicyclo- and tricylco-benzenoid compounds (67 musks and 81 nonmusks) was studied. Discriminants developed from a set of 14 molecular structure descriptors, which conveyed information about the size, shape, and electronic properties of the molecules, correctly assigned every compound of the training set into its respective category. The discriminant was successfully validated by cross-validation. To further test the predictive ability of these descriptors and the linear discriminant associated with them, an external prediction set of 15 compounds was employed. The 14 molecular descriptors that correctly classified every compound in the training set were generated for all of the compounds in the prediction set. The values of these descriptors were autoscaled using the means and standard deviations of the original training set data. Of the 15 compounds, 14 were correctly classified by the linear discriminant used for the training set. The results of this study indicate that several molecular parameters rather than a single molecular parameter are necessary to predict musk odor quality. Molecular shape was an important factor for the prediction of musk odor quality but it was by no means the only factor.

In the current study, the development of a new methodology to facilitate the design of new odorants such as musks with specialized properties is described. In a traditional framework, the introduction of a new odorant is a lengthy, costly, and laborious discovery, development, and testing process. We propose to streamline this process utilizing large existing olfactory databases available through the open scientific literature as input for a new structure/activity correlation methodology. The first step in this process is to characterize each molecule in the database by an appropriate set of descriptors. To accomplish this task, an enhanced version of Breneman's Transferable Atom Equivalent (TAE) descriptor methodology[17] can be used to create a large set of Property-Encoded Surface Translation (PEST) descriptors. We have chosen PEST descriptors to represent the problem because they have been shown to correlate with molecular shape and electronic properties as well as key modes of intermolecular interactions. Traditional QSAR methodologies, which employ fragment based descriptors, have been shown to be effective for SAR development within homologous sets of molecules but are less effective when applied to data sets containing a great deal of structural variation. In contrast to previous attempts at SAR, our use of spatially resolved electron density based PEST molecular property descriptors has removed many of the limitations brought about by the use of descriptors based on substructure fragments, molecular surface properties, or other whole molecule descriptors. Another important reason for the mixed success of past SAR efforts can be traced to the nature of the underlying modeling problem, which is often quite complex. To meet these challenges, a genetic algorithm for pattern recognition analysis has been developed that selects descriptors which create class separation in a plot of the two largest principal components of the data while simultaneously searching for features that increase the clustering of the data. The efficacy of this methodology has been evaluated using a structurally diverse database consisting of 331 macrocyclic and nitroaromatic compounds (192 musks and 139 nonmusks).

DATA SET

All compounds used in this study were taken from literature reports of chemical structure and odor quality.[18−23] A list of the compounds comprising the database is given in Table 1. The 192 macrocyclic and nitroaromatic musks are of strong, medium, weak, or of unspecified odor intensity, whereas the 139 nonmusks are odorless or have an odor other than musk. Information about odor quality and intensity is contained in the label associated with each compound. It should be emphasized that a musk compound labeled as weak, medium, or strong refers only to the change in its odor threshold, not in its odor quality.

We deliberately chose the nonmusks to be as similar in structure to the musks as possible. This not only adds the extra challenge of separating very similar structures according to odor quality but also increases our understanding of how small structural changes can affect odor quality. The structural classes present in the data set are shown in Figure 1. Natural musks, whose sources include both rare animal and plant species, are macrocycles, whereas the first synthetic musks prepared were nitrated derivatives of benzene.

MOLECULAR DESCRIPTORS

Two-dimensional representations of the compounds were encoded by drawing each compound on a graphics terminal with Chembase (Molecular Design Limited), which converted the graphical representation of the structures into computer compatible connection tables. The set of connection tables for the 331 compounds were then used to generate molecular descriptors. In addition, a three-dimensional mo-

Table 1. Compounds Comprising the Musk Database[a]

index	label	compound name
1	MSTR	8-cyclohexadecanone
2	MSTR	2(1H) benzocyclododecenone,3,4,5,6,7,8,9,10,11,12,13,14-dodecahydro
3	MSTR	cyclotetradecone, 4-methyl
4	MWEA	2(1H) benzocyclododecenone, tetradecahydro-
5	MMED	cyclopentadecanol
6	MSTR	cyclopentadecanone, 5-methyl
7	MSTR	4-cyclopentadecen-1-one
8	MSTR	2-cyclopentadecen-1-one, 3-methyl
9	MSTR	4-cyclopentadecen-1-one, (Z)-
10	MMED	17-oxabicyclo[14.1.0]heptadecane
11	MSTR	5-cyclohexadecen-1-one
12	MMED	thiacyclopentadecane
13	MMED	oxacyclotridecan-2-one
14	MMED	oxacyclotridecan-2-one, 14-methyl-
15	MSTR	oxacyclotridecan-2-one, 15-methyl-
16	MSTR	oxacyclotridec-6-en-2-one, (Z)-
17	MSTR	oxacyclotridecan-2-one, 16-methyl-
18	MSTR	1,5-dioxacyclopentadecan-2-one
19	MMED	oxacyclopentadecane-2,15-dione
20	MSTR	cyclopentadecanone, 2-hydroxy-
21	MMED	1,5-dioxacyclopentadecane-6,15-dione
22	MSTR	oxacyclohexadec-11-en-2-one
23	MSTR	1,3-dioxacycloheptadec-10-en-2-one
24	MSTR	1,4-dioxacyclohexadecane-5,16-dione
25	MOTH	2-cyclotetradecen-1-one
26	MMUS	3-cyclotetradecen-1-one
27	MOTH	1-(1-cyclododecen-1-yl)-1-acetyl-1-cyclododecene
28	MOTH	1-butanone, 1 -(1-cyclododecen-1-yl)-1-butyryl-1-cyclododecene
29	MMUS	1-cyclododecene-1-methanol,.alpha. -methyl-
30	MMUS	cyclododecanemethanol, alpha. -methyl-
31	MSTR	cycloheptadecanone
32	MMUS	9-cycloheptadecen-1-one, (Z)
33	MSTR	cyclopentadecanone
34	MWEA	cyclooctadecanone
35	MSTR	cyclohexadecanone
36	MSTR	cyclotetradecanone
37	MMUS	1,5-dioxacyclopentadecan-2,4-dione
38	MMUS	1,4-dioxacyclotetradecan-2,3-dione
39	MMUS	1,3-dioxacyclopentadecan-2-one
40	MMUS	oxacyclotridecan-2,13-dione
41	MMUS	azacycloheptadecane
42	MSTR	oxacycloheptadecan-2-one
43	MMUS	oxacycloheptadec-8-en-2-one
44	MSTR	cyclcopentadecanone, 3-methyl-
45	MWEA	cyclopentadecanone, 2-methyl-
46	MMUS	cyclopentadecanone, 4-methyl-
47	MSTR	oxacyclohexadecan-2-one
48	MMUS	1,3-dioxacycloheptadecan-2-one
49	MSTR	oxacyclopentadecan-2-one
50	MMUS	1,3-Dioxacyclohexadecan-2-one
51	MMUS	oxacyclotetradecan-2-one
52	MMUS	1,4- dioxacyclopentadecan-2,3-dione
53	MMUS	1H-cyclopentacyclododecan-1-one, tetradecahydro
54	MMUS	2(1H) benzocyclododecenone 4,5,6,7,8,9,10,11,12-dihydro
55	MOTH	cyclododecapyrimidine, 5,6,7,8,9,10,11,12,13,14 -decahdyro-N
56	MOTH	cyclododecan
57	MOTH	bicyclo [10.1.0] tridecane, 1-methoxy
58	MMUS	oxacycloletadec-9-en-2-one
59	MMUS	cyclohexadecanolid
60	MMUS	ethanone, 1-(2,5,5-trimethylcycloheptyl)-
61	MMUS	oxacyclohexadecane-2,13, dione
62	MMUS	oxacyclohexadecane-2,13, dione, 16,16, dimethyl
63	MMUS	5-cyclopentadecen-1-one (Z) and (E)
64	MMUS	cyclotetradecanon
65	MWEA	oxacyclopentadecane
66	MWEA	oxacyclohexadecane
67	MMUS	thiacyclotetradecane
68	MMUS	thiacycloheptadecane
69	MMUS	azacyclotetradecane
70	MMUS	azacyclotetradecane, 1-methyl-
71	MMUS	azacyclopentadecane
72	MMUS	azacyclopentadecane, 1-methyl-
73	MMUS	azacyclohexadecane
74	MMUS	azacyclohexadecane, 1-methyl-
75	MMUS	azacycloheptadecane, 1-methyl-

Table 1. (Continued)

index	label	compound name
76	MMUS	azacyclooctadecane
77	MMUS	azacyclooctadecane, 1-methyl-
78	MMUS	cyclotridecanone
79	MSTR	cyclohexadecanone, 3-methyl
80	MMUS	oxacyclohexadecan-6-one
81	MMUS	1,3-dioxacyclotetradecan-2-one
82	MMUS	1,3-dioxacycloheptadecane
83	MWEA	oxacyclooctadecan-2-one
84	MMUS	cyclopentadecom-1,2,3-trione
85	MMED	1,4-dioxacycloheptadecan-5-one
86	MMED	1,5-dioxacycloheptadecan-6-one
87	MMUS	1,4-dioxacyclohexadecane-2,3-dione
88	MMUS	1,4-dioxacyclotetradecane-5,14-dione
89	MMUS	1,4-dioxacyclopentadecane-5,15-dione
90	MMUS	1,5-cyclopentadecanedione
91	MWEA	2H-benzocyclotridecen-2-one, hexadecahydro
92	MWEA	cyclononadecanone
93	MMED	1,7-dioxacycloheptadecan-2-one
94	MWEA	1-oxa-5-thiacyclohexadecan-2-one
95	MMUS	5-cyclotetradecen-1-one, (Z)-
96	MMUS	1,4-dioxacycloheptadecane-5,17-dione
97	MMUS	pyridin
98	MMUS	16-azabicyclo[10.3.1]hexadeca-1(16),12,14-triene, 3-methyl-,(.+-.)-
99	MSTR	1,8-dioxacylcycloheptadecan-9-one
100	MSTR	1,7-dioxacycloheptadecan-8-one
101	MMED	1,6-dioxacycloheptadecan-7-one
102	MWEA	1-oxa-6-thiacycloheptadecan-17-one
103	MWEA	1-oxa-7-thiacycloheptadecan-18-one
104	MWEA	1-oxa-5-thiacycloheptadecan-16-one
105	MMUS	1,5-dioxacyclohexadecan-6-one
106	NOTH	cyclododecanone
107	NOTH	bicyclo [7.2.0] undec-s-en-2-ol, 2,6,10,10-tetramethyl
108	NOTH	bicyclo [7.2.0] undec-5-ene-2-acetaldehyde, 6,10,10-trimethyl
109	NOTH	cyclododecan-1,3,dioxole, 3a alpha, 4,5,6,7,8,9,10,11,12,13,13a beta-dodecahydro-2-methyl
110	NOTH	cyclododeca - 1,2 dioxane, 5,6,7,8,9,10,11,12,13,14-decahydro
111	NOTH	cyclododeca [b] furan, tetradecahydro
112	NOTH	cyclododecane, 1-methoxy-2 methyl
113	NOTH	5-cyclododecene-1-methoxy methoxy
114	NOTH	ketone, methyl, 2,6,10-trimethyl-1 cyclododecen-1yl
115	NOTH	cyclododecan[c] furan, 1,3,3a,4,5,6,7,8,9,10,11,13a-dodecahydro
116	NOTH	cyclododecyl-1,5,9-trimethy, 10-acetyl, 1,8-diene
117	NOTH	cyclodocecyl-1-ene-1,5,9-trimethyl-5,6-epoxide
118	NOTH	3-cyclooctene-1-methanol, 7-hydroxy, diformate
119	NOTH	cyclooctane (methoxy methoxy)
120	NOTH	cyclooct [e] isobenzofuran, tetradecahydro-3a-methyl
121	NOTH	6-azulenol, 2,3,3A,4,5,6,7,8-octahydro-1 methyl-4 methylene-7-(1-methylethyl)
122	NOTH	6(1H)-azulenone, 2,3,3A,7,8,8A,hexahydro-3A-methyl-1-(1methylethyl)-
123	NOTH	4-azulenemethanol, decahydro-8-methyl-2-(1-methylethenyl)-acetate
124	NOTH	cyclononanone
125	NFAI	1,3-dioxolan-2-one
126	NFAI	1,3-dioxan-2-one
127	NOTH	1,3-dioxonan-2-one
128	NOTH	1,3-dioxecan-2-one
129	NOTH	1,3-dioxacyloundecan-2-one
130	NOTH	1,3-dioxacyclotridecan-2-one
131	NOTH	1,3-dioxepane
132	NOTH	1,3-dioxocane
133	NOTH	1,3-dioxonane
134	NOTH	1,3-dioxacyclododecane
135	NOTH	1,3-dioxacyclotridecane
136	NOTH	oxacycloundecan-2-one
137	OLES	2,5-furandione, dihydro
138	OLES	2H-pyran-2,6 (3H)-dione, dihydro
139	NFAI	2,7-oxepandedione
140	NOTH	2,10-oxecanedione
141	NOTH	oxacycloundecan-2,11-dione
142	NOTH	oxacyclododecane-2,12-dione
143	OLES	cyclopentadecan-1,7-dione
144	OLES	cyclopentadecan-1,8-dione
145	NOTH	cyclodecanone
146	NOTH	cycloundecanone
147	NOTH	cyclododeane (1,1-dimethylethoxy)
148	NOTH	cyclododecane, (methoxy methyl)-
149	NOTH	cyclododecane, (ethoxymethyl)-
150	NOTH	cyclododecane, [(1-methylethoxy) methyl]-

Table 1. (Continued)

index	label	compound name
151	NOTH	cyclododecane, [2-propenyoxy) methyl]-
152	NOTH	cyclopentanone
153	NOTH	cyclohexanone
154	NOTH	cycloheptanone
155	NOTH	cyclooctanone
156	OLES	cycloeicosanone
157	OLES	cycloheneicosanone
158	NONM	2-cyclopentadecen-1-one
159	NONM	oxacyclotridecan-2-one, 13-methyl
160	NONM	1,3 dioxacyclooctadec-2-one
161	NONM	oxacyclotetradecane-2,14-dione
162	NONM	1,4-cyclopentadecanedione
163	NONM	1,6-cyclopentadecanedione
164	NONM	1H-indene, 2,3-dihydro-1,1,2,4,6-pentamethyl-5,7-dinitro
165	NONM	benzene, 1,3,5-trinitro
166	NONM	benzene, 2-methyl-1,3,5-trinitro
167	NONM	benzene, 1,3,5-trimethyl-2,4,6-trinitro
168	NONM	benzene, 2,4,dimethyl, 1,3,5-trinitro
169	NONM	benzene, 1,3-bis (1,1-dimethylethyl)-2,4-dinitro
170	NONM	benzene, 2-methoxy-1,3-bis (1-methylethyl)-5-nitro
171	NONM	benzene, 2-(1,1-dimethylethyl)-1,3,5-trinitro
172	NONM	benzene, 1,5-bis (1,1-dimethylethyl)-3, methyl-2,4-dinitro
173	NONM	benzene, 2-methoxy-4-methyl-1-(4-methylpentyl)-3,5-dinitro
174	NONM	benzene, 1-(1,1-dimethylethyl)-2-methoxy-4 methyl-3-nitro
175	NONM	benzene, 1,4-bis (1,1-dimethylethyl)-2-methoxy-3-nitro
176	NONM	benzene, 1,3-bis (1,1-dimethylethyl)-2-methoxy-4-methyl-5-nitro
177	NONM	benzaldehyde, 6-(1,1-dimethylethyl)-2-methoxy-3-nitro
178	NONM	benzene, 1-(1,1-dimethylethyl)-2methoxy-3,4,dimethyl-5-nitro
179	NONM	benzene, 1-(1,1-dimethylethyl)-2,3-dimethoxy-4-nitro
180	NONM	benzoic acid, 2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro
181	NONM	benzene, 2-butyl-4-methyl-1,3,5-trinitro
182	NONM	benzene, 2-methyl-4 (3-methylbutyl)-1,3,5-trinitro
183	NONM	benzaldehyde, 2,6-bis (1,1-dimethylethyl)-3-methoxy-5-nitro
184	NONM	benzene, 1-bromo-4 butyl-2-methyl-3,5-dinitro
185	NONM	benzamine, 4-(1,1-dimethylethyl)-2,6-dimethyl, 3,5-dinitro
186	NONM	benzene, 1-(1,1-dimethylethyl)-4-methoxy-3-methyl-2,5,6-trinitro
187	NONM	benzene, 1,3-dibromo-4-(1,1-dimethylethyl)-2,6-dimethyl, 5-nitro
188	NONM	benzaldehyde, 4-(1,1-dimethylethyl)-3-nitro
189	NONM	benzaldehyde, 4-(1,1-dimethylethyl)-3,5-dinitro
190	NONM	benzene, 2-(1,1-dimethylethyl)-4,5 dimethyl-1,3-dinitro
191	NONM	benzoic acid, 4-(1,1-dimethylethyl)-2-methyl 3,5-dinitro
192	NONM	benzene, 1(1,1-dimethylethyl)-3,4-dimethyl-6- (1-methylethyl)-2,5 dinitro
193	NONM	benzene, 1(1,1-dimethylethyl)-4 methoxy-3,5-dimethyl-2-nitro
194	NONM	benzene, 2-methoxy-1,3 dimethyl-4,5 dinitro
195	NONM	benzene, 1(1,1-dimethylethyl)-3,5-dimethyl-2,4-dinitro-6[phenylmethyl]sulfonyl
196	NONM	benzamine, 2(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro
197	NONM	benzene, 1(1,1-dimethylethyl)-3-methoxy-2,4-dinitro
198	NONM	methanone, [2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitrophenyl] phenyl-
199	NONM	benzene, 1-(1,1-dimethylethyl)-3,5-dimethyl-2,6-dinitro-4 [phenylmethyl] sulfonyl
200	NONM	benzene, 2-ethyl-5-isooctyl-4-methoxy-1,3-dinitro
201	NONM	benzene, 1-isoheptyl-2-methoxy-4-methyl-3,5-dinitro
202	NONM	benzene, 1-isooctyl-2-methoxy-4-methyl-3,5-dinitro
203	NONM	ethanone, 1- [3-(1,1-dimethylethyl)-2-methoxy-5-nitrophenyl
204	NOTH	benzene, 1,3-dibromo-2-isopropyl-5-methoxy-4-nitro
205	OLES	benzene, 1,3-bis-(1,1-dimethylethyl)-5-nitro
206	OLES	benzaldehyde, 5 (1,1-dimethylethyl)-2-methoxy-3-nitro
207	OLES	benzene, 1,5-bis- (1,1-dimethylethyl)-2 methoxy-4 methyl 3-nitro
208	OLES	ethanone, 1- [3-(1,1-dimethylethyl)-4-methoxy-5-nitrophenyl
209	OLES	benzaldehyde, 2 (1,1-dimethylethyl)-4-methoxy-5-nitro
210	OLES	benzene, 1,4-bis- (1,1-dimethylethyl)-2 methoxy-5-nitro
211	OLES	benzene, 1-(1,1-dimethylethyl)-2,5-dimethoxy-4-nitro
212	OLES	benzene, 1-(1,1-dimethylethyl)-2-methoxy-4 methyl-5-nitro
213	NOTH	benzaldehyde, 5- (1,1-dimethylethyl)-2-methyl-3-nitro
214	OLES	benzaldehyde, 2- (1,1-dimethylethyl)-4,5,6-trimethyl-3-nitro
215	OLES	benzene, 5-(1,1-dimethylethyl)-2-methoxy-1,3-dinitro
216	OLES	benzene, 1-(1,1-dimethylethyl)-4-methoxy-2-methyl 3-5-dinitro
217	OLES	benzene, 2-bromo-5- (1,1-dimethylethyl)-4-methoxy-1,3-dinitro
218	OLES	benzene, 2-butoxy-1- (1,1-dimethylethyl)-4-methyl-3,5-dinitro
219	OLES	benzene, 2-methoxy-4 methyl-1- (1-methylpropyl)-3,5-dinitro
220	OLES	benzene, 2-methoxy-4 methyl-1- (2-methylpropyl)-3,5-dinitro
221	OLES	benzene, 2-(1,1-dimethylethyl)-4-methoxy-5-methyl-1,3-dinitro
222	OLES	benzene, 1-(1,1-dimethylethyl)-5-methoxy-2,4-dintro
223	OLES	benzaldehyde, 2-(1,1-dimethylethyl)-4-methoxy-3,5-dinitro
224	OLES	phenol, 4-(1,1dimethylethyl)-2,6-diethyl-3,5-dinitro
225	OLES	benzene, 1-(1,1-dimethylethyl)-4-methoxy-3,5-dimethyl-2,6-dinitro

Table 1. (Continued)

index	label	compound name
226	OLES	benzene, 5-(1,1-dimethylethyl)-2-methyl-1,3-dintro
227	OLES	benzene, 2,5-bis- (1,1-dimethylethyl)-1,3-dinitro
228	OLES	benzene, 5-(1,1-dimethylpropyl)-2-methyl-1,3-dintro
229	OLES	benzene, 1-(1,1-dimethylethyl)-3,5-diethyl-2,4,6-trinitro
230	OLES	benzofuran, 2,3 -dihydro-3,3,6-trimethyl-5,7-dinitro
231	OLES	benzene, 1,4-bis-(1,1-dimethylethyl)-2-methoxy-3,5-dinitro
232	OLES	benzenemethanol, 4-(1,1-dimethylethyl)-2,6-dimethyl -3,5-dinitro
233	OLES	1H-indene, 6-(1,1-dimethylethyl)-2,3 dihydro-4,5,7-trinitro
234	OLES	benzaldehyde, 2,4-bis- (1,1-dimethylethyl)-5-methoxy-3-nitro
235	NONM	benzene, 2-(1,1-dimethylethyl)-4-methoxy-1,3,5-trinitro
236	NONM	benzene, 2-methoxy-1- (2-methylpropyl)-3,5 dinitro-4-(trifluoromethyl)
237	NONM	benzene, 1-(1,1-dimethylbutyl)-2-methoxy-4-methyl-3-5-dinitro
238	NONM	benzene, 1-(1,1-dimethylpentyl)-2-methoxy-4-methyl-3-5-dinitro
239	NONM	benzene, 2-methoxy-1,5-bis (1methylethyl)-3-nitro
240	NONM	benzene, 1-(1,1-dimethylpentyl)-2-methoxy-4,5-dimethyl-3-nitro
241	NONM	benzenemethanol, 4 - (1,1-dimethylethyl)-2,6-dimethyl -3,5-dinitro-, acetate (ester)
242	NONM	benzene, 1-(1,1-dimethylethyl)-3,4,5-trimethyl-2-nitro
243	NONM	benzene, 1-(1,1-dimethylethyl)-2,3,4,5-tetramethyl-6-nitro
244	NONM	1H-indene, 2,3 -dihydro-1,1,4,6-tetramethyl 5,7-dinitro
245	NONM	1H-indene, 6-ethyl-2,3 dihydro-1,1-dimethyl 5,7-dinitro
246	MMUS	benzene, 1-bromo-4 (1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro
247	MSTR	benzene, 1-azido-4 (1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro
248	MSTR	ethanone, 1-[2(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitrophenyl]
249	MMUS	benzaldehyde, 2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro
250	MMUS	benzonitrile, 4-(1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro
251	MMUS	benzene, 1-(1,1-dimethylethyl)-2-fluoro-3,5-dimethyl-2,6-dinitro
252	MMUS	benzene, 1-(1,1-dimethylethyl)-4-fluoro-3,5-dimethyl-2,6-dinitro
253	MMUS	ethanone, 1-(4-methoxy-2-methyl -3,5-dinitrophenyl)
254	MMUS	ethanone, 1-[4(1,1-dimethylethyl)-2,3,6-trimethyl-5-nitrophenyl]
255	MSTR	benzene, 1-(1,1-dimethylethyl)-3,5-dimethyl-2,4,6-trinitro
256	MMUS	benzene, 1-bromo-3- (1,1-dimethylethyl)-5-methyl-2,4, 6-trinitro
257	MMUS	benzene, 2-(1,1-dimethylethyl)-4-methyl-1,3,5-trinitro
258	MMUS	benzene, 1-chloro-3- (1,1-dimethylethyl)-5-methyl-2,4,6-trinitro
259	MWEA	benzene, 1,3-bis (1,1-dimethylethyl)-5-methyl-2,4,6-trinitro
260	MMUS	benzene, 3-(1,1-dimethylethyl)-1-fluoro-5-methyl-2,4,6-trinitro
261	MWEA	benzene, 2-hexyl-4-methyl-1,3,5- trinitro
262	MMUS	benzene, 2-methyl-4- (1-methylethyl)-1,3,5-trinitro
263	MSTR	benzene, 1-(1,1-dimethylethyl)-2-ethoxy-3,5-dinitro
264	MSTR	benzene, 1-(1,1-dimethylethyl)-2-ethoxy-4 ethyl 3,5-dinitro
265	MSTR	benzene, 1-(1,1-dimethylpropyl)-4 ethyl-2-methoxy- 3,5-dinitro
266	MSTR	benzene, 1,3-bis (1,1-dimethylethyl)-4-methoxy-6-methyl-2,5-dinitro
267	MSTR	1H-indene, 2-ethyl, 2,3 -dihydro-1,1,3,3,5-pentamethyl 4,6-dinitro
268	MSTR	silane, (3,5-dimethyl-2,4,6-trinitrophenyl) trimethyl-
269	MMUS	benzene, 1,3-bis (1,1-dimethylethyl)-2-methoxy-5-nitro
270	MMUS	benzene, 1-(1,1-dimethylethyl)-2-methoxy-3-nitro
271	MMUS	benzene, 2-methoxy-4-methyl-1,3,5-trinitro
272	MMUS	benzaldehyde, 3-(1,1-dimethylethyl)-2-methoxy-5-nitro
273	MMUS	benzene, 2-bromo-4- (1,1-dimethylethyl)-1,3,5-trinitro
274	MMUS	benzene, 1-(1,1-dimethylethyl)-2,4-dimethoxy-5,6-dimethyl-3-nitro
275	MMUS	benzene, 2-methyl-4- (2-methylpropyl)-1,3,5-trinitro
276	MMUS	benzene, 5-(1,1-dimethylethyl)-2-ethyl-1,3-dinitro-
277	MMUS	ethanone, 1-(2-butyl-3-methyl-4,6-dinitrophenyl)-
278	MMUS	benzene, 1-(1,1-dimethylethyl)-2-methoxy-3,5-dimethyl-4,6-dinitro
279	MMUS	benzene, 2-(1,1-dimethylpropyl)-4-methyl-3- (1-methylethyl)-1,5-dinitro
280	MMUS	benzene, 1-(1,1-dimethylethyl)-3-methyl-2,4-dinitro
281	MMUS	benzoyl chloride, 6-(1,1-dimethylethyl)-3-methyl-2,4-dinitro-
282	MMUS	benzene, 1,3-dibromo-2- (1,1-dimethylethyl)-5-methoxy-4-nitro
283	MMUS	benzonitrile, 2-(1,1-dimethylethyl)-4,5-dimethyl-3-nitro-
284	MMUS	benzene, 1-(1,1-dimethylethyl)-3-methoxy-5-methyl-2,4,6-trinitro
285	MMUS	benzene, 2-(1,1-dimethylethyl)-4-methyl-3- (1-methylethyl) 1,5-dinitro
286	MMUS	benzene, 1-(1,1-dimethylpropyl)-2,4-dimethoxy-3,5-dinitro-
287	MWEA	benzene, 2,4-bis- (1,1-dimethylethyl)-1-nitro
288	MSTR	ethanone, 1-[4-(1,1-dimethylpropyl)-2,6-dimethyl-3,5-dinitrophenyl]
289	MSTR	benzene, 1-(1,1-dimethylpropyl)-3,4,5-trimethyl-2,6-dinitro-
290	MMUS	benzene, 1-(1,1-dimethylethyl)-2-methoxy-3,5-dinitro-
291	MMUS	benzene, 4-(1,1-dimethylethyl)-6-methyl-1,3-dinitro
292	MSTR	benzene, 6-(1,1-dimethylethyl)-3-ethyl-1-methoxy-2,4-dinitro
293	MWEA	benzene, 6-(1,1-dimethylethyl)-1-methoxy-3- (1-methylethyl)-2-4-dinitro-
294	MSTR	benzene, 6-(1,1-dimethylethyl)-1-ethoxy-3-methyl-2,4-dinitro-
295	MWEA	benzene, 6-(1,1-dimethylethyl)-3-methyl-1- (1-methyl ethoxy)-2,4-dinitro-
296	MMUS	benzaldehyde, 4-(1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro-
297	MMED	benzoic acid, 2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro-methyl ester
298	MWEA	benzene, 5-(1,1-dimethylpropyl)-2-ethyl-1,3-dinitro-
299	MWEA	benzene, 2-ethyl-1,3-dinitro-
300	MSTR	benzene, 2-(1,1-dimethylpropyl)-4,6-dimethyl-1,3,5-trinitro-

Table 1. (Continued)

index	label	compound name
301	MSTR	1H-indene, 3-ethyl, 2,3 -dihydro-1,1,3,5-tetramethyl 4,6-dinitro
302	MMUS	naphthalene, 1,2,3,4- tetrahydro-1,1,4,4-tetramethyl-6-(1methylethyl)-5,7-dinitro
303	MMUS	naphthalene, 6-*tert*-butyl-1,2,3,4- tetrahydro-1,1,4,4-tetramethyl-5,7-dinitro
304	MMUS	naphthalene, 1,2,3,4-tetrahydro-7-isopropyl-1,1,2,4,4-pentamethyl-6,8-dinitro
305	MMUS	naphthalene, 1,2,3,4-tetrahydro 1,1,4-trimethyl-6,8-dinitro
306	MWEA	1-propanone, 1-[2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitrophenyl]-
307	MMUS	benzene, 1,3-dibromo-5-(1,1-dimethylethyl)-2-methoxy-6-methyl-2-nitro
308	MSTR	1H-indene, 2,3 -dihydro-1,1,3,3,5-pentamethyl 4,6-dinitro
309	MSTR	benzene, 2-(1,1-dimethylethyl)-4,5,6-trimethyl-1,3-dinitro-
310	MSTR	ethanone, 1-[4-(1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitrophenyl]-
311	MMUS	benzene, 1-(1,1-dimethylethyl)-2-methoxy-4-methyl-3,5-dinitro-
312	MMUS	ethanone, 1-[6-(1,1-dimethylethyl)-3-ethyl-2,4-dinitrophenyl]-
313	MMUS	benzene, 1-bromo-6-(1,1-dimethylethyl)-3-methoxy-2,4-dinitro-
314	MMUS	benzaldehyde, 5-(1,1-dimethylethyl)-4-methoxy-3-nitro-
315	MMUS	benzaldehyde, 5-(1,1-dimethylethyl)-4-ethoxy-3-nitro-
316	MWEA	benzene,4,6-bis- (1,1-dimethylethyl)-1-methoxy-2-nitro
317	MSTR	benzene,6-(1,1-dimethylethyl)-1,3-dimethoxy-2,4-dinitro
318	MSTR	benzene,6-(1,1-dimethylpropyl)-1-methoxy-3-methyl-2,4-dinitro-
319	MMUS	benzonitrile,2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro-
320	MMUS	benzene,1-bromo-2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro
321	MWEA	benzoyl chloride, 2-(1,1-dimethylethyl)-4,6-dimethyl-3,5-dinitro-
322	MWEA	benzoyl chloride, 4-(1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro-
323	MWEA	benzoic acid, 4-(1,1-dimethylethyl)-2,6-dimethyl-3,5-dinitro,-methylester
324	MSTR	benzene,2-(1,1-dimethylethyl)-4-ethyl-6-methyl-1,3,5-trinitro
325	MMUS	1H-indene, 2,3 -dihydro-1,1,2,3,3,5-hexamethyl 4,6-dinitro
326	MMUS	1H-indene, 5-ethyl- 2,3-dihydro-1,1,3,3 -tetramethyl 4,6-dinitro
327	MMUS	1H-inden-5-ol, 2-ethyl- 2,3-dihydro-1,1,3,3-tetramethyl 4,6-dinitro
328	MMUS	1H-indene, 5-ethyl- 2,3-dihydro-1,1,2,3,3-pentamethyl 4,6-dinitro
329	MMUS	1H-indene, 3-ethyl- 2,3-dihydro-1,1,2,3,3-pentamethyl 4,6-dinitro
330	MWEA	benzene, 1,5-bis-(1,1-dimethylethyl)-3-methyl-2-nitro-
331	NONM	benzenesulfonic acid, 4-butyl-2-methyl-3, 5-dinitro

[a] MSTR − strong musk; MMED − medium musk; MWEA − weak musk; MMUS − musk of unspecified odor strength; MOTH − musk that has a secondary odor note that is not musk; OLES − odorless; NOTH − specified odor other than musk; NONM − nonmusk.

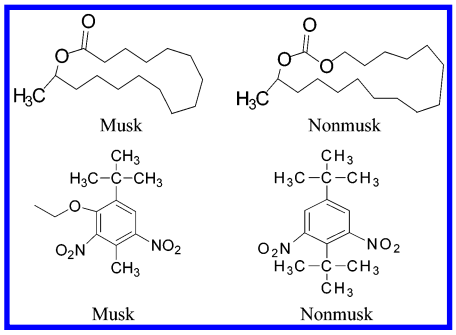

Figure 1. Two very strong musks and two odorless compounds representing the two major structural classes of compounds found in the data set.

lecular model was generated for each compound in the database by a molecular mechanics model building routine that utilized the CHARMM force field contained in the modeling program Quanta (Molecular Simulations). Traditional molecular property descriptors, including connectivity based topological 2D descriptors, physicochemical property descriptors, and shape independent 3D molecular features were computed via the MOE program (Chemical Computing Group, Montreal, Canada). TAE molecular surface property reconstructions were generated using Convert2001.[24] The Property Encoded Surface Translator (PEST) algorithm[25] was used to generate wavelet and hybrid shape/property descriptors. Molecular electron density properties for all compounds in the study were represented using TAE surface histogram descriptors, wavelet coefficient descriptors, and hybrid shape/property descriptors.

Molecular descriptors were computed individually for each structural group of musks, allowing for easy separation and modeling of the structural classes. Descriptors based on experimental data were not used in this study for two reasons. First, it would be extremely difficult to obtain this type of data from the literature for a large set of compounds. Second, it would not be possible to use the SAR developed in this study as a screening tool to identify new musks since the compounds would have to be synthesized in order to obtain the necessary data.

ELECTRON DENSITY DERIVED DESCRIPTORS

To effectively characterize the potential odor quality of any molecule, an appropriate set of molecular descriptors must be utilized. While many types of descriptors exist in the literature, the present work emphasizes the use of electron density derived descriptors of three general types—TAE or molecular surface property histogram descriptors, surface property wavelet coefficient descriptors (WCD), and PEST surface property hybrid descriptors. The underlying methodology involved in TAE, WCD, and PEST descriptor generation relies on the hypothesis that a causative relationship exists between observed odor properties and the distribution of certain molecular electronic properties as sampled on molecular van der Waals surfaces. An additional hypothesis for PEST shape/property hybrid descriptor validation is that the spatial arrangement of surface electronic properties contains pertinent chemical information. Both of these hypotheses have been previously validated[26] in studies involving biological and nonbiological molecular behavior.

Quantum chemical descriptors offer an attractive alternative to traditional fragment based descriptors by expressing a more accurate and detailed description of the electronic

Table 2. TAE Atomic Electronic Surface Properties

EP	electrostatic potential
Del(Rho) •N	electron density gradient normal to 0.002 e/au^3 electron density isosurface
G	electronic kinetic energy density $G = (-(\hbar/4m)\int\{\nabla\psi^* \cdot \nabla\psi\}\, d\tau)$
K	electronic kinetic energy density $K = -(\hbar/4m)\int\{\psi^*\nabla^2\psi + \psi\nabla^2\psi^*\}\, d\tau)$
Del(K) •N	gradient of K electronic kinetic energy density normal to surface
Del(G) •N	gradient of B electronic kinetic energy density normal to surface
Fuk	Fukui F^+ function scalar value
Lapl	Laplacian of the electron density $\nabla^2\rho$
BNP	bare nuclear potential $BNP_{(j)} = \sum_{i=1}^{n} q_i/r_{ij}$
PIP	local average ionization potential $PIP(r) = \sum_i \rho_i(r)\lvert\epsilon_i\rvert/\rho(r)$

and geometric properties and the interaction between them. However, even with rapid advances in computer architecture and the anticipated continued growth in computational power, a direct calculation of the properties of large molecules at a high level of theory is prohibitive. Bader's quantum theory of Atoms in Molecules (AIM) provides a framework for reconstructing large complicated molecules from a number of small electron density fragments while still achieving a good approximation of the properties of the intact molecules.[27,28] In AIM Theory, the electron density of a molecule can be partitioned into distinct electron density basins (the regions of space occupied by the corresponding atoms), each containing an atomic nucleus. These electron density fragments are essentially bounded by surfaces of zero net flux in the electron density, which correspond to the steepest descent pathways in rho from each bond critical point. An atomic property (A) can then be expressed as the integral of a corresponding property density $\rho_A(r)$ over an atomic basin:

$$A(\Omega) = \int_\Omega d\tau \rho_A(r), \quad \text{where}$$

$$\rho_A(r) = (N/2)\int d\tau'\{\psi^*\hat{A}\psi + (\hat{A}\psi)^*\psi\} \quad (1)$$

$A(\Omega)$ is the integrated property associated with the atom, and $d\tau_\Omega$ represents all space within the basin bound by zero-flux surfaces. Each property density $\rho(r)$ may be represented as an expectation value for an operator A. The expression for property $\rho_A(r)$ is shown as a property of wave function ψ as illustrated on the right of eq 1. These atomic properties possess a high degree of transferability from the electronic environment in one molecule to another molecule with a similar environment. Consequently, the properties of a functional group or whole molecule can be obtained by adding these atomic properties together.

Based on AIM Theory, Breneman introduced the concept of "Transferable Atom Equivalent".[29] A TAE is an atomic electron density fragment bounded by interatomic zero flux surfaces and an extended isodensity surface that approximates the condensed phase van der Waals surface. TAE fragments carry 10 atomic charge density derived properties (see Table 2) that are precomputed from small molecules using ab initio wave functions at the 6-31+G* level of theory. TAE electron density reconstructions provide not only molecular electron densities but also electronic kinetic energy densities and local average ionization potentials as well as other first and second derivative properties of the electron density. The distributions of these electronic properties computed on 0.002-e/au^3 electronic density isosurfaces may be characterized as molecular property descriptors in several ways. TAE histogram descriptors can be produced by recording the distribution of these properties as surface histograms that quantify the molecular surface areas with specific ranges of each property value. In addition to these histogram descriptors, property extrema, average values, and standard deviations of the property distributions (in some cases with separate σ values for positive and negative portions of the range) are also included in the TAE descriptor set.

The TAE library consists of a set of precalculated atomic fragments structured in a form that allows the atomic fragments involved in the new molecule to be rapidly retrieved. The RECON program reads the atomic connectivity information of the molecule and assigns the closest fragment match from the TAE library to each atom based on atom type, hybridization, and structural environment. By summing up the corresponding atomic properties of the constituent fragments, we can obtain a large set of electron density based TAE descriptors. These descriptors provide information about basicity, hydrophobicity, hydrogen bonding capacity, and polarity as well as molecular polarizability.

Surface property distributions can also be characterized by the use of discrete wavelet transforms. Wavelet coefficient descriptors (WCDs) produced through TAE reconstructions are also additive at the atomic level and convey more chemical property information as a result of the data compression methods used to generate them. The TAE property encoded surface can also be subjected to Zauhar "Shape Signature" ray tracing approach[30] to generate PEST descriptors. A ray is initialized with a random location and direction within the molecular surface and reflected throughout inside the electron density isosurface until the molecular surface is adequately sampled. Molecular shape information is obtained by recording the ray path information including segment lengths, reflection angles, and property values at each point of incidence. Path information (segment length and point of incidence values) can be summarized into 2D histograms to obtain a surface shape profile. For a single electronic property, a 2D histogram having the distribution of distance (x-axis) versus the associated property value (y-axis) can give a characteristic distribution (z-axis) based on the overall shape and property value of the molecule. Such a 2D histogram can be created for every surface property. Each bin of the 2D histogram becomes a hybrid shape/property descriptor.

PATTERN RECOGNITION ANALYSIS

For pattern recognition analysis, each compound was initially represented by 896 computer generated molecular descriptors that were derived from connection tables or from three-dimensional models of the compounds. Before any given descriptor was entered into the study, it was checked to see whether it had the same value for every compound in

the training set. A descriptor was eliminated from consideration if it was invariant. Prior to pattern recognition analysis, all descriptors were autoscaled to zero mean and unit standard deviation to alleviate problems arising from scaling.

The basic premise underlying the approach to pattern recognition used in this study is that all data analysis methods work well when the problem is simple. By identifying the appropriate features, a "hard" problem can be reduced to a "simple" one. Therefore, our goal is feature selection to increase the signal-to-noise ratio of the data by discarding molecular descriptors that are not characteristic of the various odorant classes in the data. To ensure identification of all relevant descriptors, it is best that a multivariate approach to feature selection be employed. The approach must also take into account the existence of redundancies in the data.

In this study, a genetic algorithm (GA) for pattern recognition analysis was used to identify molecular descriptors from which discriminating relationships could be found. The pattern recognition GA selects descriptors that increase clustering while simultaneously searching for descriptors that optimize the separation of the classes in a plot of the two or three largest principal components of the data. The principal component analysis routine embedded in the fitness function of the GA acts as an information filter, significantly reducing the size of the search space, since it restricts the search to features whose principal component plots show clustering on the basis of class. In addition, the algorithm focuses on those classes and or samples that are difficult to classify as it trains using a form of boosting to modify the class and sample weights. Samples that consistently classify correctly are not as heavily weighted as those samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network.

The fitness function of the GA emulates human pattern recognition through machine learning to score the principal component plots and thereby identify descriptors that optimize the separation of the odorant classes in a plot of the two or three largest principal components of the data. To facilitate the tracking and scoring of the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see eqs 2 and 3) where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set). $SW_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights for the objects comprising a particular class sum to a value equal to the class weight of the class in question.

$$\mathrm{CW(c)} = 100\frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}}\mathrm{CW(c)}} \tag{2}$$

$$\mathrm{SW(s)} = \mathrm{CW(c)}\frac{\mathrm{SW(s)}}{\sum_{\mathrm{s}\in\mathrm{c}}\mathrm{SW(s)}} \tag{3}$$

Each principal component plot generated for each feature subset after it has been extracted from its chromosome is scored using the *K*-nearest neighbor classification algorithm.[31] For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is taken of the point's K_c nearest neighbors. For the most rigorous classification, K_c equals the number of samples in the class to which the point belongs. (K_c often has a different value for each class.) The number of K_c nearest neighbors with the same class label as the sample point in question, the so-called sample hit count, SHC(s), is computed ($0 \le \mathrm{SHC(s)} \le K_c$) for each sample. It is then a simple matter to score a principal component plot (see eq 4). First, the contribution to the overall fitness by each observation or compound in class 1 is computed, with the scores of the compounds comprising the class summed to yield the contribution by this class to the overall fitness. This same calculation is repeated for the other classes with the scores from each class summed to yield the overall fitness, *F(d).*

$$F(d) = \sum_{\mathrm{c}}\sum_{\mathrm{s}\in\mathrm{c}}\frac{1}{K_{\mathrm{c}}} \times \mathrm{SHC(s)} \times \mathrm{SW(s)} \tag{4}$$

The fitness function of the GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations. (Boosting the weights is referred to as adjusting the internal parameters in the block diagram of the pattern recognition GA.) To boost, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c over all feature subsets produced in a particular generation (see eq 5), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a class (see eq 6). ϕ in eq 5 is the number of chromosomes in the population, and AVG in eq 6 refers to the average or mean value. During each generation, class and sample weights are adjusted by a perceptron (see eqs 7 and 8) with the momentum, P, set by the user. (g + 1 is the current generation, whereas g is the previous generation.) Classes with a lower class hit rate are boosted more heavily than those classes that score well.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i\mathrm{(s)}}{K_{\mathrm{c}}} \tag{5}$$

$$\mathrm{CHR}_{\mathrm{g}}\mathrm{(c)} = \mathrm{AVG}(\mathrm{SHR}_{\mathrm{g}}\mathrm{(s)}\colon \forall_{\mathrm{s}\in\mathrm{c}}) \tag{6}$$

$$\mathrm{CW}_{\mathrm{g+1}}\mathrm{(s)} = \mathrm{CW}_{\mathrm{g}}\mathrm{(s)} + \mathrm{P}(1 - \mathrm{CHR}_{\mathrm{g}}\mathrm{(s)}) \tag{7}$$

$$\mathrm{SW}_{\mathrm{g+1}}\mathrm{(s)} = \mathrm{SW}_{\mathrm{g}}\mathrm{(s)} + \mathrm{P}(1 - \mathrm{SHR}_{\mathrm{g}}\mathrm{(s)}) \tag{8}$$

Boosting is crucial for the successful operation of the pattern recognition GA because it modifies the fitness landscape by adjusting the values of the class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the GA will change as the population is evolving toward a solution.

RESULTS AND DISCUSSION

The first step in the study was to apply principal component analysis to the data. Principal component analysis is a method for transforming the original measurement variables into new uncorrelated variables called principal

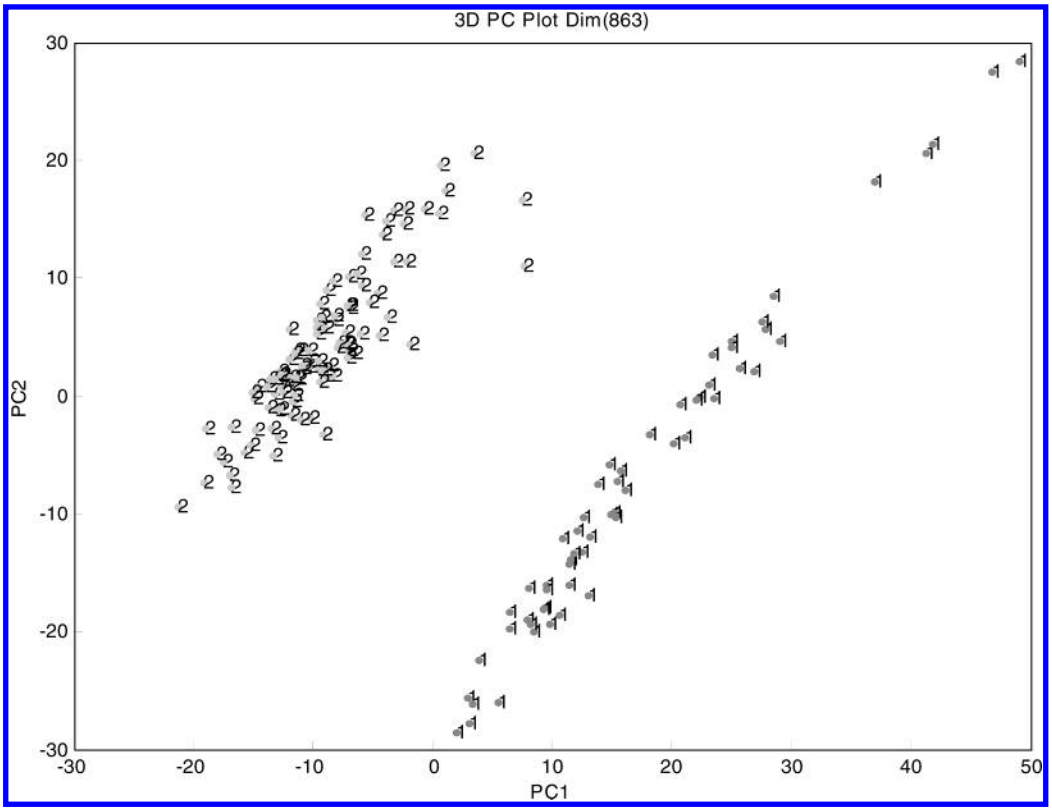

Figure 2. A plot of the two largest principal components of the 163 macrocycles and 863 TAE and MOE descriptors. 1's are nonmusks, and 2's are musks. The plane defined by the two largest principal components accounts for 45% of the total cumulative variance of the data.

components. Each principal component is a linear combination of the original measurement variables. Using this procedure is analogous to finding a new coordinate system that is better at conveying information present in the data than axes defined by the original measurement variables. The coordinate system is linked to variance. Often, only two or three principal components are necessary to explain all the information present in a data set if there are a large number of interrelated measurement variables. Using principal component analysis, dimensionality reduction, classification of samples, and identification of clusters in high dimensional data are possible.

MACROCYCLES

In Figure 2, the results of a principal component mapping experiment are shown for the 163 macrocycles (index numbers 1−163 in Table 1) and the 863 TAE and MOE descriptors. The 1's in Figure 2 are nonmusks, and the 2's are musks. It is evident that macrocyclic musks are well separated from macrocyclic nonmusks in the principal component plot. Because this projection is made without the use of information about the class assignments of the compounds, the resulting separation is therefore a strong indication of real differences in the molecular descriptor profile of macrocyclic musks and nonmusks.

The pattern recognition GA was used to find a smaller set of molecular descriptors from which a discriminating relationship could be developed for the macrocyclic musks. We identified these molecular descriptors by sampling key descriptor subsets, scoring their principal component plots, and tracking those classes and or compounds that were difficult to classify. After 100 generations, the pattern recognition GA identified nine TAE derived molecular descriptors whose principal component plot showed clustering on the basis of odor quality (see Figure 3). Because the musks are well separated from the nonmusks on the first principal component, one must conclude that musk odor activity can be accurately modeled by TAE derived descriptors. The nine TAE derived molecular descriptors that were selected by the pattern recognition GA are listed in Table 3. DGNW7 and DGNW22, which characterize the rate of change in the G kinetic energy density normal to and away from the surface of the molecule, describe weak bonding interactions. DKNH14, DKNW4, and DKNW13, which describe the rate of change in the K kinetic energy density normal to and away from the surface of the molecule, are correlated to hydrophobicity and polarizability. GW27 and KW27, which are wavelet descriptors derived from the G and K kinetic energy reconstructions normal to and away from the surface of the molecule, describe hydrogen-bonding interactions. DKNB30 and BNPB31, which are PEST descriptors, suggest that molecular shape is important. These PEST descriptors capture more of the interior volume (i.e., local shape) information as opposed to conformational information. This is in contrast to the Grid-Independent descriptors of Cruciani.[32]

The SAR of macrocyclic musks is well documented in the literature. Fifteen- and 16-member ring compounds with the appropriate functional group have odors in which the musk odor modality predominates. As the ring size increases, musk odor character is retained, but an animal note appears and the odor intensity decreases rapidly until 20- or 21-member ring compounds, which are odorless. The musk odor modality first appears in 13-member rings as an impure note.

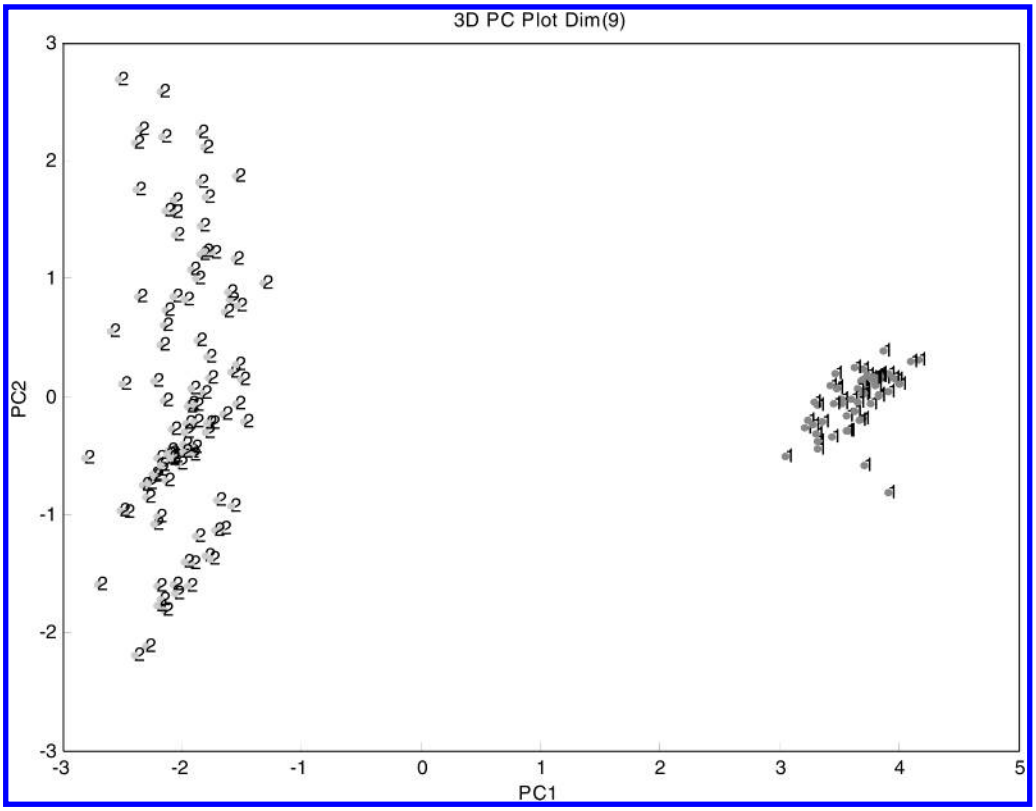

Figure 3. A plot of the two largest principal components of the 163 macrocycles and nine TAE derived descriptors identified by the pattern recognition GA. 1 = nonmusk, and 2 = musk. The plane defined by the two largest principal components accounts for 52% of the total cumulative variance.

Table 3. Final Subset of Descriptors for Macrocyclic Musks

DGNW7	scale wavelet descriptor ("W") describing the rate of change of the G kinetic energy density normal to and away from the surface
DGNW22	detail coefficient wavelet descriptors ("W") describing the rate of change of the G kinetic energy density normal to and away from the surface
DKNH14	rate of change of the K kinetic energy density normal to and away from the surface— The "H" defines this descriptor as a histogram descriptor, and the "14" means that its value is from the 14th bin. Evidently, larger values of DKN are important for musk odor quality.
DKNW4	scale wavelet descriptor emphasizing the importance of changes in the K kinetic energy
DKNW13	coefficient wavelet descriptor emphasizing the importance of changes in the K kinetic energy
DKNB30	shape descriptor— The "3" indicates that rays of average size are important, and the "0 " indicates that very low DKN values are important.
GW27	coefficient wavelet descriptor emphasizing the importance of the electronic kinetic energy density G
KW27	coefficient wavelet descriptor emphasizing the importance of the electronic kinetic energy density K
BNPB31	shape descriptor developed from the bare nuclear potential—The "3" indicates that rays of average size are important, and the "1" indicates that relatively low BNP values are important.

Stretched conformations are believed to play a more important role in the generation of musk odor than circular ones. The introduction of a second functional group is generally injurious to musk odor unless the distance between the functional groups is small enough to render cooperation between them. The 163 macrocycles that were selected for this study capture these trends, and the principal component map of the nine descriptors identified by the pattern recognition GA has effectively captured this information by way of a graphically oriented structure activity correlation. As for Figure 3, it is important to note that both the musks and nonmusks have approximately the same covariance matrix, which does not appear to be the situation after examination of the eigenvector plot because of a mapping error that has occurred when projecting the musks onto the principal component plane developed from the nine TAE derived descriptors.

NITROAROMATIC MUSKS

Figure 4 shows the results of a principal component mapping experiment for the 168 nitroaromatics, which were divided into a training set of 151 compounds and a prediction set of 17 compounds chosen by random lot (see Table 4). The compounds in the predictions set were projected onto a principal component map developed from the 151 training set compounds and 859 molecular descriptors. The 1's are nonmusks, the 2's are musks, M represents musks from the prediction set, and N represents nonmusks from the prediction set. From an examination of this plot, it is evident that musks are well separated from the nonmusks. Furthermore, every compound in the prediction set is projected into a region of the map occupied by compounds possessing the same odorant label. Because this projection of the data is developed from a large number of descriptors, the resulting

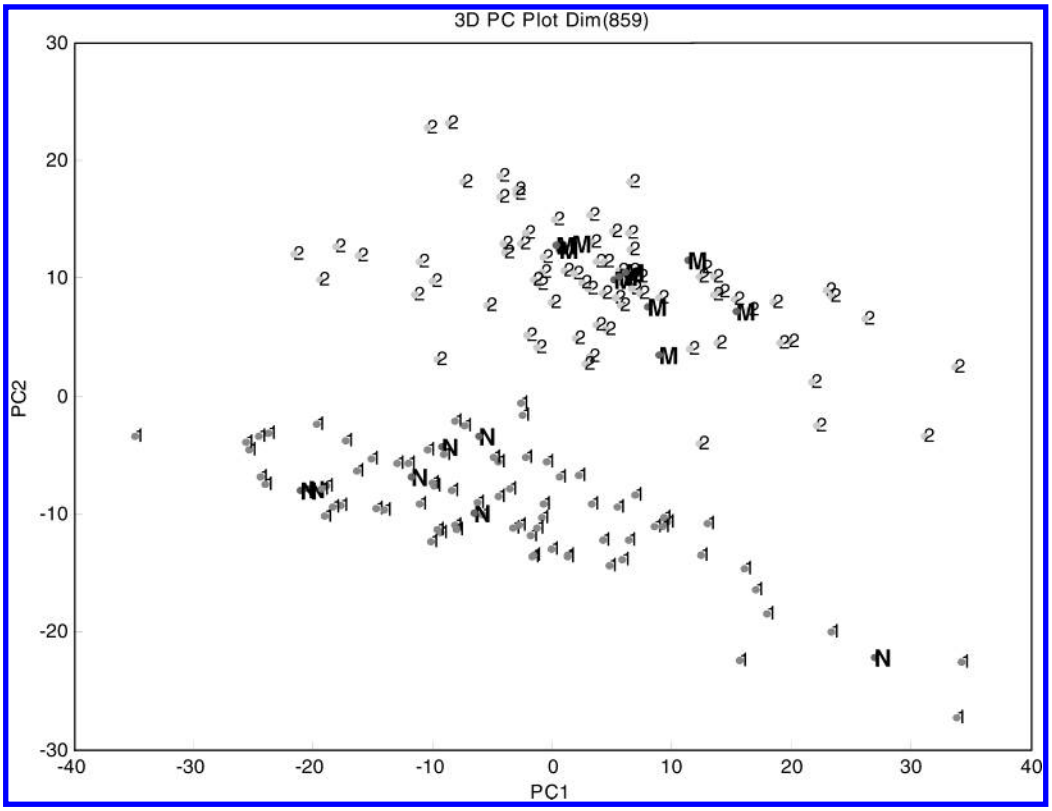

Figure 4. A plot of the two largest principal components of the 151 aromatic nitro musks and 859 TAE and MOE descriptors. 1 = nonmusk and 2 = musk. Predictions set compounds (M = musk and N = nonmusk) are shown projected onto the principal component map developed from the training set data. The plane defined by the two largest principal components accounts for 40% of the total cumulative variance.

Table 4. Aromatic Nitro Musk Data Set

Training Set (index numbers/Table 1): 164, 165, 167−192, 194−201, 203−230, 232−237, 239, 241−244, 246−263, 266−285, 287−289, 291−294, 296−299, 301−311, 313−316, 318, 320, 321, 323−331
Prediction Set (index numbers/Table 1): 166, 193, 202, 231, 238, 240, 245, 264, 265, 286, 290, 295, 300, 312, 317, 319, 322

separation is therefore a strong indication of real differences in the molecular descriptor profile of aromatic nitro musks and nitroaromatic nonmusks. Evidently, the molecular descriptors used were able to delineate the complex substitution pattern of aromatic nitro musks, which has confounded previous investigators.

The pattern recognition GA was again used to identify the molecular descriptors most characteristic of musk odor. Key descriptors were identified by sampling key feature subsets and scoring their principal component plots, while tracking those classes and/or samples that were difficult to classify. The pattern recognition GA identified six TAE derived descriptors whose principal component plot showed clustering of the compounds in the training set on the basis of odor. When the compounds in the predictions set were projected onto the principal component map developed from the 151 training set compounds and six molecular descriptors, the prediction set compounds again were projected onto a region of the map occupied by compounds with the same class label (see Figure 5). The six molecular descriptors identified by the pattern recognition GA are listed in Table 5. Three are shape descriptors and three are wavelet descriptors. DGNW14, a wavelet descriptor, which characterizes the rate of change in the G kinetic energy density normal to and away from the surface of the molecule, describes weak bonding interactions with the receptor. DKNW18, a wavelet descriptor, which describes the rate of change in the K kinetic energy density normal to and away from the surface of the molecule, is correlated to hydrophobicity and polarizability. DGNB32 and DGNB34 are shape descriptors derived from ray traces of the G kinetic energy normal to and away from the surface of the molecule. EPW3, which is a wavelet descriptor derived from the molecule's electrostatic potential, is correlated to the solvation energy of the molecule. EPB01 is a shape descriptor derived from ray traces of the molecule's electrostatic surface potential.

Many of the changes in the chemical structure of the aromatic nitro musks studied involved a shift of a methyl group, which would not appear to be significant when inspecting their structural formulas. However, these types of changes can have a profound impact on intermolecular interactions involving an odorant and a condensed phase. For example, the shift of a methyl group from a position adjacent to a secondary hydroxyl group to a position one methylene group away can profoundly change the polarity of the compound on Carbowax 20M by 100 Kovats index units because the OH group has become more accessible for intermolecular interactions. Evidently, the six TAE derived descriptors identified by the pattern recognition GA contain information about changes in the intermolecular interactions of the odorants.

The SAR of aromatic nitro musks is not well understood because of the complex substitution pattern and the varied polyfunctional character of the nitro group. Aromatic nitro musks have highly impure and informationally complex

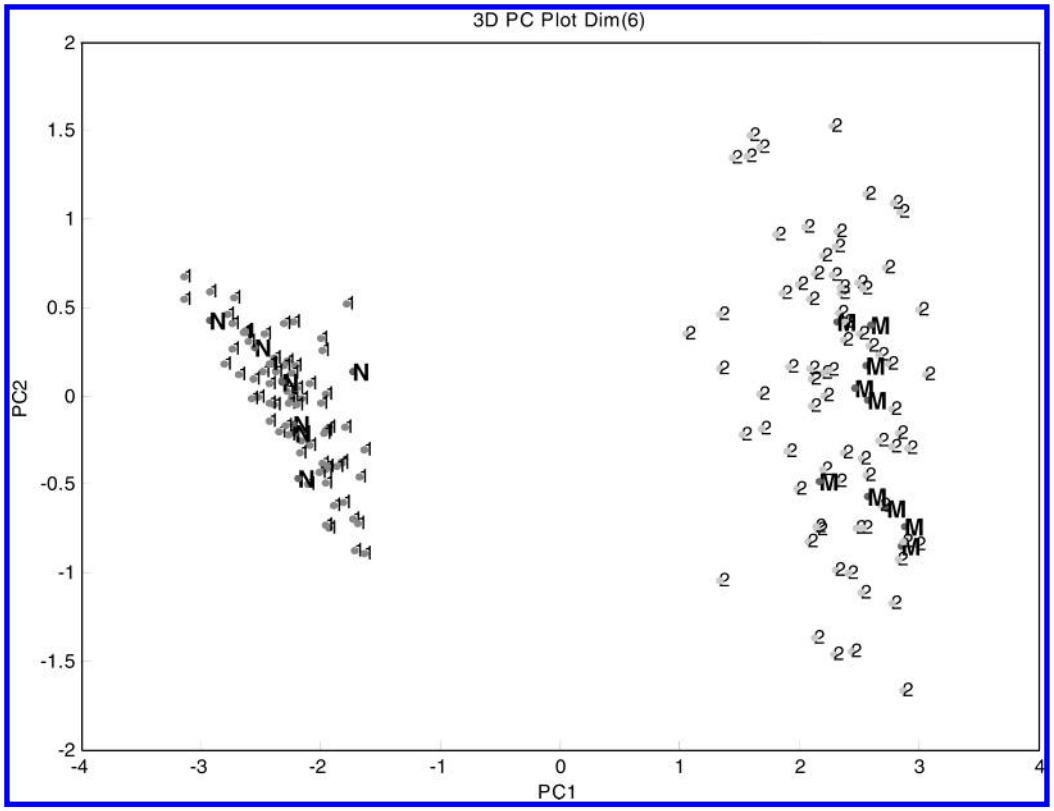

Figure 5. A plot of the two largest principal components of the 151 aromatic nitro musks and six TAE derived descriptors identified by the pattern recognition GA. 1 = nonmusk, and 2 = musk. Prediction set compounds (M = musk and N = nonmusk) are shown projected onto the principal component map developed from the six TAE derived descriptors. The plane defined by the two largest principal components accounts for 47% of the total cumulative variance.

Table 5. Final Subset of Descriptors for Nitrated Musks

DGNW14	detail coefficient wavelet descriptor ("W") describing the rate of change of the G kinetic energy density normal to and away from the surface
DGNB32	shape descriptor—The "3" states that rays of average size are important, and the "2 " states that relatively low DGN values are important.
DGNB34	shape descriptor—The "3" indicates that rays of average size are important, and the "4 " indicates that relatively high DGN values are important.
DKNW18	detail coefficient wavelet descriptor emphasizing the importance of the change in the K kinetic energy
EPW3	scale coefficient wavelet descriptor describing the electrostatic potential on the surface of the molecule
EPB01	shape descriptor—The "0" indicates that short rays are important and the "1" indicates that low electrostatic potential values are important.

odors. Nevertheless, we could separate aromatic nitro musks from nonmusks using a single principal component developed from six TAE derived descriptors. One can therefore conclude that musk odor activity of aromatic nitro musks can be accurately modeled by TAE derived descriptors. In our opinion, this constitutes an important step forward in the study of olfactory relationships since it has been demonstrated via macrocylic and aromatic nitro musks that TAE derived descriptors convey significant information about molecular interactions important in olfaction.

JOINT STUDY

Because aromatic nitro musks are so different in odor quality from other musks, many scientists and perfumers[33] have considered the possibility of a different response mechanism for them. However, Beets and co-workers[34] have shown that nitrogen free and aromatic nitro musks share some similarities in their substitution pattern. To better understand the relationship between structure and odor quality for musks, the same SAR methodology previously used to study the nitro and macrocyclic musks was applied to the entire database. This larger study was undertaken for several reasons. There is the opportunity to resolve conflicting views on factors influencing musk odor quality thereby bringing some order into this field by uncovering structural features that are common to all musks. As a result, it may be possible to validate hypotheses on substitution patterns of aromatic nitro musks initially proposed by Beets to rationalize their anomalous behavior. This study would also serve, as a test to assess whether TAE derived descriptors are capable of spanning structural manifolds. Our understanding of how small changes in chemical structure affect odor quality would be broadened since this issue would be considered in the context of a more structurally diverse data set.

Figure 6 shows a principal component plot of the 331 compounds and the 871 TAE derived and MOE descriptors. The 1's are the macrocyclic nonmusks, 2's are the aromatic nitro nonmusks, 3's are the macrocyclic musks, and 4's are aromatic nitro musks. It is evident from the plot that most of the information captured by the principal component map is about chemical structure since the macrocycles are well separated from the nitro aromatics on the first principal component. To identify molecular descriptors correlated with musk odor quality, it was again necessary to use the pattern recognition GA, which sampled key feature subsets, scored

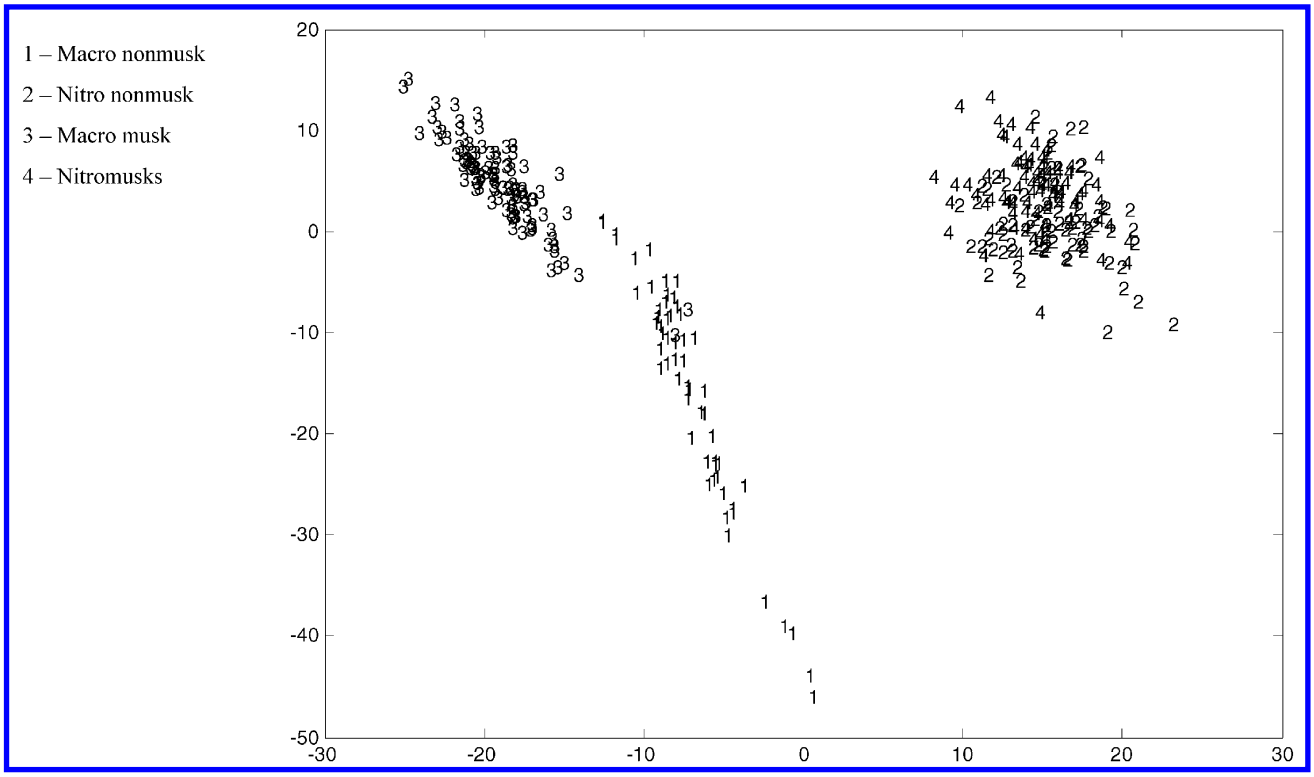

Figure 6. A plot of the two largest principal component plot of the 331 compounds and the 871 TAE and MOE descriptors comprising the musk database. 1 = macrocyclic nonmusk, 2 = aromatic nitro nonmusks, 3 = macrocyclic musks, and 4 = aromatic nitro musks. The plane defined by the two largest principal components accounts for 35% of the total cumulative variance.

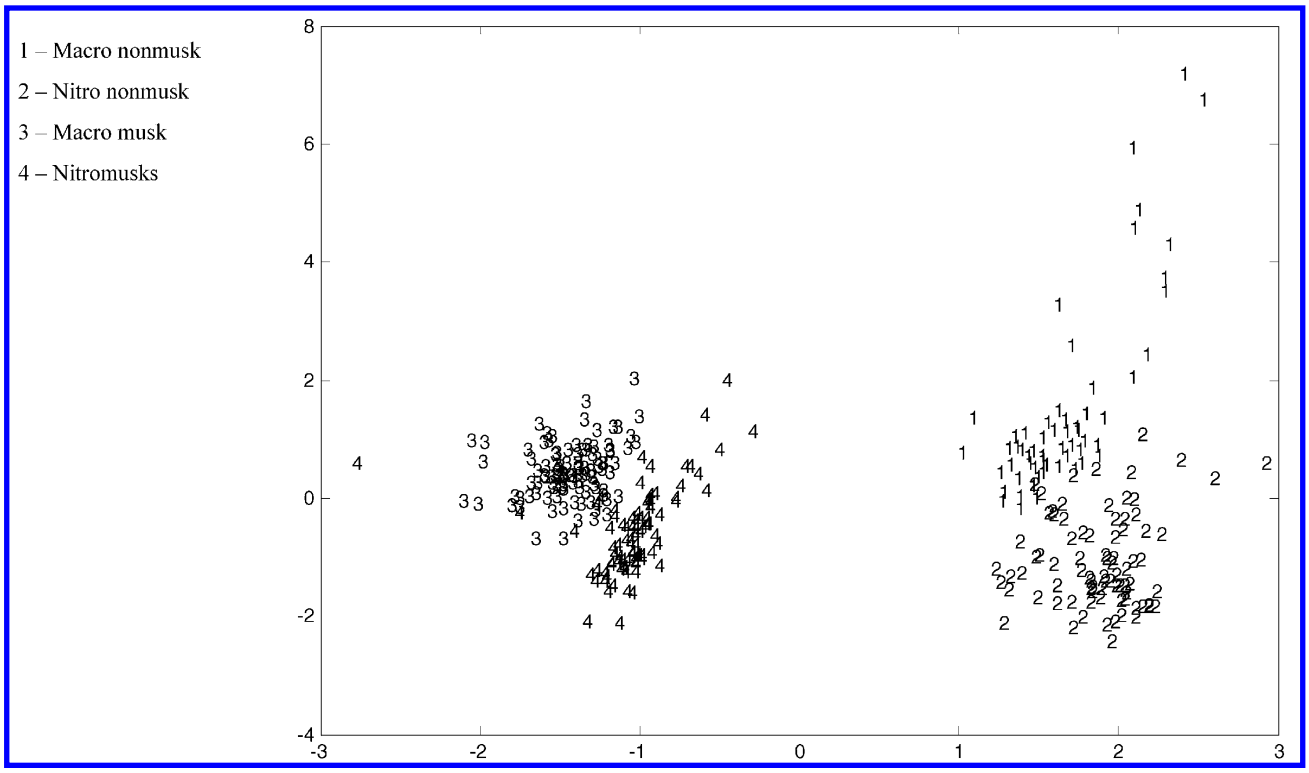

Figure 7. A plot of the two largest principal component plot of the 331 compounds and 16 TAE derived descriptors identified by the pattern recognition GA. 1 = macrocyclic nonmusk, 2 = aromatic nitro nonmusks, 3 = macrocyclic musks, and 4 = aromatic nitro musks. The plane defined by the two largest principal components accounts for 57% of the total cumulative variance.

their principal component plots, while tracking those classes and/or samples that were difficult to classify. The pattern recognition GA identified 16 TAE derived descriptors whose principal component plot (see Figure 7) showed clustering of the compounds on the basis of odor quality. Molecular descriptors that were identified by the pattern recognition GA are listed in Table 6. Most of the 16 descriptors identified by the pattern recognition GA convey information about intermolecular interactions, which suggests their importance in defining musk odor quality. DGNH8 represents the rate

Table 6. Final Subset of Descriptors for the Musk Database

DGNH8	rate of change of the G kinetic energy density normal to and away from the surface—The "H" defines this descriptor as a histogram descriptor, and the "8" means that its value is from the 8th bin. Evidently, large values of DGN are important for musk odor quality.
DGNW7	scale wavelet descriptor ("W") describing the same basic value type as DGNH8
DGNW21	detail coefficient wavelet descriptor ("W") describing the same basic value type as DGNH8
DGNB52	shape descriptor—The "5" indicates that long rays are important and the "2" indicates that low DGN values are important.
LAPLB30	The Laplacian is the second derivative of the electronic energy distribution. It is a shape descriptor with "3" meaning that intermediate length rays are represented, and "0" meaning that small property values are represented.
DKNMAX	maximum value of DKN—It is similar to DGNMAX except that it uses K kinetic energy density, which is often complementary to the G kinetic energy density.
DKNW6	scale wavelet descriptor emphasizing the importance of changes in the K kinetic energy
DKNB55	shape descriptor representing relatively long rays between relatively high values of DKN
DRNH6	This descriptor represents the rate of fall off of the electron density. The "H" defines this descriptor as a histogram descriptor, and the "6" means that its value is from the 6th bin. This type of descriptor is often highly correlated to DGN.
DRNW26	a detail wavelet descriptor for DRN
PIPH8	PIP is the local average ionization potential. The "8" represents a bin containing lower than average values for PIP
PIPW18	detail coefficient wavelet descriptor for PIP
PIPW29	detail coefficient wavelet descriptor for PIP
BNPW4	scale coefficient wavelet descriptor of the bare nuclear potential—It is suspected of describing polar and hydrogen bonding interactions.
BNPW9	scale coefficient wavelet descriptor of the bare nuclear potential—It is suspected of describing polar and hydrogen bonding interactions.
ANGLEB52	special shape descriptor—For long rays (5), the angle (2) is less than average.

of change of the G kinetic energy density normal to and away from the surface of the molecule. DGNW7 and DGNW21 are wavelet descriptors describing the same basic value. All three descriptors are correlated to weak bonding interactions and probably describe some facet of the interaction between the musk and the receptor. DGNB52, a shape descriptor, can be interpreted to mean that it is probably crucial for areas with a low rate of change in the G kinetic energy to be relatively far apart in the molecule. LAPLB30, another shape descriptor, is the second derivative of the electronic energy distribution and is important in characterizing donor/acceptor relationships. DKNMAX, which is the maximum value for the rate of change of the K kinetic energy density, is complementary to G kinetic energy density descriptors. This particular DKN descriptor is correlated to hydrophobicity and polarizability. DKNW6 is a wavelet descriptor emphasizing the same properties encoded in DKNMAX, whereas DKNB55 is a shape descriptor developed from the rate of change of the K kinetic energy density. Evidently, long distances between relatively high values of the rate of change in the K kinetic energy density is crucial for musk odor quality. Both DRNH6 and DRNW26 represent the rate of fall off of the electron density and are highly correlated to the four DG descriptors previously discussed. PIPH8, PIPW18, and PIPW29 are descriptors that convey information about the local ionization potential of the molecule. BNPW4 and BNPW9 are so-called bare nuclear potential descriptors, which probably describe interactions involving polar and hydrogen bonding. ANGLEB52 is a special type of shape descriptor. Molecules that are curved and do not have sharp turns are favored by this descriptor.

Figure 7 shows that nitrated and nitro-free musks have common structural features that can be used to differentiate them from nonmusks. We consider this to be a significant result. Fragrance chemists have long sought to find the overlap between nitrated and nitro free musks in terms of the structural features that a compound must possess in order to evoke a musky odor. According to the most recent theory of olfaction, compounds interact with multiple olfactory receptors rather than individual ones. Because the olfactory receptors are not very selective toward specific odorants, identification of an odor is based on a distinct pattern of responses. It is, therefore, plausible that nitro musks such as "musk ketone" and "musk ambrette" long used in the fragrance industry because of their fixatives properties generate a response pattern, which in some measure overlaps with the response pattern generated by nitro-free musks. This would explain the SAR results obtained in this study and would also validate a viewpoint long held by perfumers that nitrated musks fit better in the musk odor category than in any other odor category even though they elicit an odor response that is markedly different from nitro free musks.

REFERENCES AND NOTES

(1) Wright, R. H. Odor and Molecular Vibration: Neural Coding of Olfactory Information. *J. Theor. Biol.* **1977**, *64*, 473−502.

(2) Turin, L. A Spectroscopic Mechanism for Primary Olfactory Reception. *Chem. Senses* **1996**, *21*, 773−791.

(3) Beets, M. G. J. *Molecular Structure and Organoleptic Quality*; Macmillan: New York, 1957.

(4) Theimer, E.; Davies, J. T. Olfaction, Musk Odor, and Molecular Properties. *J. Agric. Food. Chem.* **1967**, *15*(1), 6−14.

(5) Amoore, J. E. *Molecular Basis of Odor*; Charles C. Thomas: Springfield, IL, 1970.

(6) Dravnieks, A.; Laffort, P. Physicochemical Basis of Quantitative and Qualitative Odor Discrimination in Humans. In *Olfaction and Taste*; Schneider, D., Ed.; Wissens-Verlag-MBH: Stuttgart, FRG, 1972; pp 142−148.

(7) Axel, R. The Molecular Logic of Smell. *Sci. Am.* October **1995**, 154−159.

(8) Brugger, W. E.; Jurs, P. C. Extraction of Important Molecular Features of Musk Compounds Using Pattern Recognition Techniques. *J. Agric. Food Chem.* **1977**, *25*(5), 1158−1164.

(9) Klopman, G.; Ptscelintsev, D. Application of the Computer Automated Structure Evaluation Methodology to a QSAR Study of Chemoreception. Aromatic Musky Odorants. *J. Agric. Food Chem.* **1992**, *40*, 2244−2251.

(10) Kier, L. B.; Di Paolo, T.; Hall, L. H. Structure Activity Studies on Odor Molecules Using Molecular Connectivity. *J. Theor. Biol*. **1977**, *67*, 585−596.
(11) Greenberg, M. J. Dependence of Odor Intensity on the Hydrophobic Properties of Molecules. A Quantitative Structure Odor Intensity Relationship. *J. Agric. Food Chem*. **1979**, *27*, 347−352.
(12) Wolkowski, Z. W.; Moccatti, D.; Heymans, F.; Godfroid, J. J. A Quantitative Structure−Activity Approach to Chemoreception, Importance of Lipophilic Properties. *J. Theor. Biol*. **1977**, *66*, 181−193.
(13) Rossiter, K. J. Structure-Odor Relationships. *Chem. Rev*. **1996**, *96*, 3201−3240.
(14) Frater, G.; Bajgrowicz, J. A.; Kraft, P. Fragrance Chemistry. *Tetrahedron* **1998**, *54*, 7633−7703.
(15) Cherqaoui, D.; Esseffar, M'H.; Villemin, D.; Cense, J.-M.; Chastrette, M.; Zakarya, D. Structure-Musk Odor Relationship of Tetralin and Indan Compounds Using Neural Networks. *New J. Chem.* **1998**, *22*, 839−843.
(16) Narvaez, J. N.; Lavine, B. K.; Jurs, P. C. Structure−Activity Studies of Musk Odorants Using Pattern Recognition: Bicyclo and Tricyclo-Benzenoids. *Chem. Senses* **1986**, *11*(1), 145−156.
(17) Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. *Comput. Chem*. **1995**, *19*(3), 161−172.
(18) Beets, M. G. J. *Structure Activity Relationships in Human Chemoreception*; Applied Science Publishers: London, 1978.
(19) Wood, T. F. Chemistry of the Aromatic Musks. *Givaudanian*; Clifton, NJ, 1970; pp 1−37.
(20) Bersuker, I. B.; Dimoglo, A. S.; Gorbachov, M. Yu; Vlad, P. F.; Pesaro, M. Origin of Musk Fragrance Activity: The Electron-Topological Approach. *New J. Chem*. **1991**, *15*, 307−320.
(21) MacLeod, A. J. In *Olfaction in Mammals, Symposia of the Zoological Society of London*; Stoddard, D. M., Ed.; Academic Press: New York, 1980; No. 45, p 15.
(22) *Fragrance Chemistry, The Science of the Sense of Smell*; Theimer, E. T., Ed.; Academic Press: New York, 1982.
(23) Ham, C. L.; Jurs, P. C. Structure Activity Studies of Musk Odorants Using Pattern Recognition: Monocyclic Nitrobenzenes. *Chem. Senses* **1985**, *10*(4), 491−502.
(24) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*(6), 1347−1357.
(25) Breneman, C.; Bennett, Bi, J.; Song, M.; Embrechts, M. New Electron Density-Derived Descriptors and Machine Learning Techniques for Computational ADME and Molecular Design. MidAtlantic Computational Chemistry Meeting, Princeton University, Princeton, NJ, April 17, 2002.
(26) Breneman, C. M.; Rhem, M. QSPR Analysis of HPLC Column Capacity Factors for a Set of High-Energy Materials Using Electronic van der Waals Surface Property Descriptors Computed by Transferable Atom Equivalent Method. *J. Comput. Chem*. **1997**, *18*(2), 182−197.
(27) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford Press: Oxford, UK, 1990.
(28) Bader, R. F. W.; Becker, P. Transferability of Atomic Properties and the Theorem of Hohenberg and Kohn. *Chem. Phys. Lett.* **1988**, *148*, 452−458.
(29) Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D. Transferable Atom Equivalent Multi-centered Expansion Method. *J. Comput. Chem*. **2003**, *24*(4), 512−529.
(30) Zauhar, R. J.; Welsh, W. J. Application of the Shape Signatures Approach to Ligand-and Receptor Based Drug Design, Abstracts of Papers of the American Chemical Society 220: 70-COMP Part 1 August 20, 2000.
(31) James, M. *Classification*; John Wiley & Sons: New York, 1992.
(32) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-Independent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **2000**, *43*(17), 3233−3243.
(33) Tenahsi, R. Odor and Molecular Structure. In *Gustation and Olfaction*; Ohloff, G., Thomas, A. F., Eds.; Academic Press: London, 1971.
(34) Beets, M. G. J. Odor, Taste, and Molecular Structure. In *The Contribution of Chemistry to Food Supplies*; Morton, I., Rhodes, D. N., Eds.; Butterworth and Co.: London, 1974; pp 99−152.

CI030016J