Prediction of the Aroma Quality and the Threshold Values of Some Pyrazines Using Artificial Neural Networks[§]

Bettina Wailzer,[†] Johanna Klocker,[†] Gerhard Buchbauer,[‡] Gerhard Ecker,[‡] and Peter Wolschann*[,†]

Institute of Theoretical Chemistry and Molecular Structural Biology, University of Vienna, Währinger Strasse 17, A-1090 Vienna, Austria, and Institute of Pharmaceutical Chemistry, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria

Received December 5, 2000

An artificial neural network is used to predict both the classification of aroma compounds and their flavor impression threshold values for a series of pyrazines. The classification set consists of 98 compounds (32 green, 43 bell-pepper, and 23 nutty smelling pyrazines), and the regression sets consist of 24 green and 37 bell-pepper odorous pyrazines. The best classification of the three aroma impressions (93.7%) is obtained by using a multilayer perceptron network architecture. To predict the threshold values of bell-pepper fragrance, a standard Pearson *R* correlation coefficient of 0.936 for the training set, 0.912 for the verification set, and 0.926 for the test set is received with two hidden layers consisting of two and one neurons. The network for the threshold prediction of the class of green-smelling pyrazines with one hidden layer containing three neurons turns out to be the best with a standard Pearson *R* correlation coefficient of 0.859 for the training, 0.918 for the verification, and 0.948 for the test set. These good correlations show that artificial neural networks are versatile tools for the classification of aroma compounds.

Introduction

The relationship between the molecular structure of flavor compounds and the intensity as well as the quality of their aroma impression has received more and more interest within the past years. This led to a better understanding of the physicochemical mechanism of both flavor and odor perception. Odorant binding proteins (OBPs), which are identified as members of the lipocalin superfamily, are necessary for the transportation of the aroma compounds from the air to the olfactory receptors through the aqueous barrier of the mucus. Alternatively they might remove the odorant molecules from the receptor after the transduction of the olfactory signal.[1] This superfamily includes several secretory proteins often interacting specifically with small, mainly hydrophobic ligands. They were identified in bovine olfactory mucosa, in mucosa of rats, mice, rabbits, and pigs, and in other animals.[2−5] The three-dimensional structures have been determined by X-ray investigations on bovine and porcine OBP,[6] and some ideas about the binding site for odor compounds have been proposed as well. For other receptor proteins studied, only the amino acid sequences are known. Both large differences in primary structures and the existence of several OBPs in the same animal species suggest different binding sites for odorants and pheromones.[5]

Because no extended information from the 3D structures of the receptor binding sites is available, other ligand-based methods have to be applied. Quantitative structure−activity relationships (QSAR) techniques are widespread and rather successful methods in modern drug design, and therefore, these methods should also give more insight into aroma chemistry. Nevertheless, in the case of complex relationships, conventional QSAR methods often lead to unsatisfactory results because of nonlinear relationships within the data set. In some cases, explicit nonlinear functions, such as the bilinear model for log *P*/log(potency) dependencies, have to be used on a trial and error basis.[7] Moreover, if membrane-bound receptors are involved, the biological activity often is the result of both membrane interaction and receptor binding, which also may lead to nonlinear dependencies. One possibility of overcoming the difficulties of such nonlinearities in QSAR and 3D-QSAR studies is the use of artificial neural networks (ANNs), which gained increasing interest in the field of drug design.[8] After a proper learning procedure, ANNs should be able to “recognize” basic correlations in a given data set and to predict, for example, physicochemical properties and pharmacological activities. Several applications of ANNs in structure−activity relationships of aroma and odor compounds have been already described. Chastrette et al.[9] have investigated 79 nitrobenzene derivatives with musk fragrance, using a multilayer back-propagation neural network. In another study Chastrette et al.[10,11] have used the same method, with both a three-layer neural network and a multilayer neural network, to investigate a series of tetralins and indans. Furthermore, Zakarya et al.[12] tested a classification of camphor odor compounds by means of ANNs with a back-propagation algorithm and *K*th nearest neighbor. They also studied the relationship between sandalwood odor and molecular structures of organic compounds, in particular cyclohexyl-, norbornyl-, campholenyl-, and Decalin derivatives, with a three-layer back-propagation neural network.[13] Moreover, Chas-

[§] Dedicated to Prof. W. Fleischhacker on the occasion of his 70th anniversary.

* To whom correspondence should be addressed. E-mail: Karl.Peter.Wolschann@univie.ac.at. Phone: 0043 1 4277 52772. Fax: 0043 1 4277 9527.

[†] Institute of Theoretical Chemistry and Molecular Structural Biology.

[‡] Institute of Pharmaceutical Chemistry.

10.1021/jm001129m CCC: $20.00
Published on Web 07/18/2001

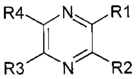


Figure 1. General structure of compounds with R1, R2, R3, R4.

trette and El Aidi [14] established a study on the classification of pyrazines and pyridines of bell-pepper aroma impression using ANNs. They used Charton's steric hindrance descriptor and group electronegativity values to distinguish between bell-pepper and non-bell-pepper aroma impressions by coding the odor with a binary variable. The results showed that a simple description of the substituents could provide enough information for the ANN to learn structure–odor rules. In extensions of our investigations on structure–flavor relationships of pyrazine-derived aroma compounds,[15] nonlinear structure–flavor relationships of a series of pyrazine-based flavor molecules are analyzed by ANNs. These aroma compounds show a broad spectrum of flavor impressions ranging from earthy, nutty, roasted, and green to bell-pepper and woody.[16,23–35] In particular, we investigated structure–flavor relationships of pyrazines with bell-pepper, with a green and nutty flavor. These rather sensitive and significant differences of the aromas of the pyrazines are a consequence of the modification of the length and the polarity of the side chains and their relative position at the heteroaromatic ring. Both the threshold values of the aroma impressions and the aroma qualities of a large number of pyrazine derivatives are considered for the development of predictive models by ANNs. It can be shown that the obtained models are satisfactory in both statistical significance and predictive ability.

Compounds and Aroma Classification

A total of 98 pyrazine derivatives (32 with green aroma, 43 with bell-pepper aroma, and 23 with nutty aroma) are selected from the literature in order to investigate those parameters that are of importance for the distinction of the characteristic flavor (Table 1).[16–18] The structures are superimposed so that all substituents containing a heteroatom should occupy position R1 (see Figure 1). If there are only alkyl groups present, the longest alkyl chain is placed at position R2. By use of these conventions for position numbering, there remain four structures that show only methyl groups or hydrogens as substituents. In these cases the methyl-containing substituent is laid on position R1. A total of 24 green and 37 bell-pepper smelling pyrazines are chosen to predict their threshold values and to compare them with the experimental values depicted in Table 1. As a quantitative measure for the odor impression, log(1/*c*) values are used, where *c* is the detection threshold value of the aroma compound dissolved in water, given in ppm, divided by the molecular mass of the molecule. The threshold value is defined as the olfactory detection threshold, that is, the ability of a test person to distinguish between water with and without aroma. This should be clearly differentiated from the recognition threshold, i.e., from the ability of a testing person to identify a distinct aroma.

Descriptors

The 3D structures of the compounds are built by the Hyperchem 5.0 software.[19] The resulting geometries are subsequently optimized by an ab initio method (Hartree–Fock on 3-21G level) implemented in the Gaussian 98 program.[20] For the obtained structures the following molecular properties are calculated, using TSAR 3.21 (Tools for Structure–Activity Relationships) software:[21] steric descriptors, e.g., molecular surface, both for the whole compound and for the four substituents, Verloop parameters, atom counting for carbon atoms of substituents R1, R2, and R3 and for oxygen atoms of substituent R1, as well as electronic descriptors (Hartree–Fock derived dipole moments and point charges on the atoms of the heterocycle and the first atom of the substituents). Additionally the sum of electrotopological indices is calculated.

The ANN analyses are performed with the TRAJAN software package.[22] Generally, feed-forward networks are used throughout the study. These networks have a characteristic layered architecture, the so-called multilayer perceptron (MLP) network architecture. They consist of one or more hidden layers between the input and output layers. This architecture enables very broad flexibility and thus allows solutions with a broad range of problems. Training a feed-forward network is an iterative process that involves repeating the presentation of the training set (containing compounds with known target outputs) to the network. After each presentation the network parameters (or weights) are adjusted so that the network's total error for all patterns in the set (as measured by an appropriate error function) is progressively reduced. This type of training is known as supervised learning. Several different algorithms for adjusting the network weights have been developed in the past. We use back-propagation as a training algorithm, which is derived from the oldest and simplest of the classical optimization techniques, the steepest descent algorithm. To have an independent check on the training progress, a subset of the compounds is reserved and not actually used in the back-propagation algorithm (verification set). It is used to track the network's error performance, to identify the best network, and to stop training. Additionally, a test set is generated that is not used in training at all and is designed to give an independent assessment of the network's performance when an entire network design procedure is completed. Statistics are separately calculated for each of the three subsets of the data set. Without using a verification set, a network with a large number of weights and a modest amount of compounds for training tends to overfit. In this case, the data are memorized rather than analyzed and the trained network usually shows low predictivity. This is in contrast to a good generalization, which is the ability of a network to perform well in classification and prediction of previously not considered data.

Results

1. Classification of the Aroma Impression of Pyrazines. For classification of the aroma impression a nominal output variable is used. This output correlates a distinct aroma quality with molecular structures and properties. The classification is performed via three output neurons by checking the output unit activation levels against two thresholds: the accept

Table 1. Structure, Threshold Values (X = No Threshold Value), and Aroma Impression (1 = Green, 2 = Nutty, and 3 = Bell-Pepper) of Pyrazines

compd	R1	R2	R3	R4	log(1/*c*)	quality	lit.
1	N(CH3)2	H	H	CH2CH(CH3)2	1.554	1	27
2	OC4H9	H	H	H	X	1	29
3	OC6H5	H	CH(CH3)2	H	2.790	1	24
4	SC2H5	H	(CH2)2CH(CH3)C2H5	H	3.629	1	24
5	N(CH3)2	CH3	H	H	2.882	1	27
6	OCH3	CH3	CH2CH(CH3)2	H	6.000	1	24
7	OCH3	CH3	CH2CH(CH3)C2H5	H	7.288	1	25
8	OCH3	CH3	CH2CH2CH(CH3)C2H5	H	5.239	1	24
9	OC2H5	CH3	CH(CH3)C2H5	H	4.209	1	24
10	OC2H5	CH3	CH2CH(CH3)2	H	4.084	1	24
11	OC2H5	CH3	CH2CH(CH3)C2H5	H	4.540	1	24
12	OC2H5	CH3	(CH2)2CH(CH3)C2H5	H	4.869	1	24
13	OC6H5	CH3	CH2CH(CH3)2	H	2.879	1	24
14	OC6H5	CH3	(CH2)2CH(CH3)C2H5	H	3.285	1	24
15	SCH3	CH3	CH2CH(CH3)2	H	4.116	1	24
16	SCH3	CH3	CH2CH(CH3)C3H7	H	4.271	1	25
17	SC2H5	CH3	CH2CH(CH3)C2H5	H	4.953	1	25
18	OCH3	COH(CH3)2	CH3	H	X	1	27
19	OCH3	COH(CH3)2	H	CH3	X	1	27
20	OCH3	COCH3	H	CH3	X	1	27
21	OCH3	COCH3	OCH3	CH3	X	1	27
22	H	C2H5	H	CH3	2.452	1	33
23	C2H5	C2H5	H	H	5.134	1	23
24	H	CH(CH3)2	CH3	CH3	3.882	1	24
25	H	C4H9	H	H	2.532	1	16
26	H	CH2CH(CH3)2	H	H	2.532	1	26
27	SCH3	CH2CH(CH3)2	H	H	5.742	1	24
28	H	C5H11	H	H	4.477	1	24
29	H	C5H11	CH3	CH3	3.296	1	25
30	OCH3	C5H11	H	H	6.954	1	16
31	CH3	(CH2)2CH(CH3)2	CH3	H	X	1	28
32	OCH3	C7H15	H	H	6.903	1	16
33	CH3	H	CH3	H	X	2	33
34	OCH3	H	H	H	X	2	16
35	OCH3	H	H	CH3	X	2	16
36	OC2H5	H	H	H	X	2	16
37	SCH3	H	H	H	X	2	16
38	SCH3	H	H	CH3	X	2	33
39	SC2H5	H	H	H	X	2	16
40	CH3	CH3	H	H	X	2	33
41	CH3	CH3	CH3	H	X	2	33
42	CH3	CH3	CH3	CH3	X	2	30
43	NHCH3	CH3	H	H	X	2	27
44	OCH3	CH3	H	H	X	2	16
45	OCH3	CH3	H	CH3	X	2	32
46	OC2H5	CH3	H	H	X	2	16
47	SCH3	CH3	H	H	X	2	16
48	SC2H5	CH3	H	H	X	2	16
49	H	C2H5	H	H	X	2	16
50	H	C2H5	CH3	CH3	X	2	25
51	CH3	C2H5	H	CH3	X	2	23
52	CH3	C2H5	CH3	H	X	2	31
53	SC2H5	C2H5	H	H	X	2	16
54	H	CH2CH(CH3)2	CH3	CH3	X	2	24
55	SC6H5	C8H17	H	H	X	2	16
56	CH3	C3H7	H	H	3.356	3	33
57	OCH3	C3H7	H	H	6.103	3	16
58	SCH3	C3H7	H	H	5.226	3	33
59	CH3	CH(CH3)2	H	H	3.930	3	34
60	OCH3	CH(CH3)2	H	H	6.802	3	16
61	OCH3	CH(CH3)2	H	CH3	6.567	3	27
62	OCH3	CH(CH3)2	CH3	H	6.521	3	27
63	OCH3	CH(CH3)2	OCH3	CH3	3.447	3	27
64	OCH3	CH(CH3)2	CH3	OCH3	2.894	3	27
65	OCH3	CH(CH3)2	OCH3	CH(CH3)2	2.524	3	27
66	SCH3	CH(CH3)2	H	H	6.553	3	33
67	OCH3	C4H9	H	H	6.521	3	24
68	SC2H5	C4H9	H	H	4.690	3	24
69	CH3	CH2CH(CH3)2	H	H	3.062	3	27
70	OCH3	CH2CH(CH3)2	H	H	7.472	3	26
71	OCH3	CH2CH(CH3)2	H	CH3	5.840	3	24
72	OCH3	CH2CH(CH3)2	CH3	H	4.840	3	24
73	OCH3	CH2CH(CH3)2	CH3	CH3	2.790	3	26
74	OCH3	CH(CH3)C2H5	H	H	6.618	3	16
75	OC2H5	C5H11	H	H	6.385	3	16

Table 1 (Continued)

compd	R1	R2	R3	R4	log(1/*c*)	quality	lit.
76	SCH3	C5H11	H	H	6.213	3	16
77	SC2H5	C5H11	H	H	5.322	3	16
78	OCH3	(CH2)2CH(CH3)2	H	H	7.456	3	27
79	OCH3	CH2CH(CH3)C2H5	H	H	7.176	3	33
80	OCH3	(CH2)3CH=CH2	H	H	6.773	3	33
81	OCH3	(CH2)2CH=CHCH3 (E)	H	H	6.101	3	33
82	OCH3	(CH2)2CH=CHCH3 (Z)	H	H	5.516	3	16
83	OCH3	C6H13	H	H	6.443	3	33
84	OCH3	(CH2)3CH(CH3)2	H	H	7.510	3	33
85	OCH3	CH2CH(CH3)C3H7	H	H	7.385	3	16
86	OCH3	C8H17	H	H	6.568	3	16
87	OC2H5	C8H17	H	H	5.072	3	16
88	SCH3	C8H17	H	H	5.532	3	16
89	SC2H5	C8H17	H	H	5.101	3	16
90	OCH3	C10H21	H	H	3.796	3	16
91	OC2H5	C10H21	H	H	3.644	3	16
92	OCH3	CH3	OCH3	CH3	2.970	3	35
93	OCH3	C2H5	H	H	X	3	18
94	OCH3	CH(CH3)C3H7	H	H	X	3	18
95	OCH3	(CH2)6CH(CH3)2	H	H	X	3	18
96	OCH3	CH2CH(CH3)C6H13	H	H	X	3	18
97	OCH3	CH2CH(CH3)2	H	CH2CH(CH3)2	X	3	18
98	OC2H5	CH2CH(CH3)2	H	H	X	3	18

threshold and the reject threshold. If the output is above the accept threshold, the case is positively classified, if it is below the reject value, the case is negatively classified. If the output is between the accept and the reject threshold the classification is unknown. Inspection of outliers shows that elimination of two green-smelling compounds (compounds . **22** and **24**) could give a better prediction of the whole set. Comparison of the aroma impressions of these two pyrazines points out that their odor quality is additionally slightly sweet. Furthermore, there are no structural similar compounds in the data set. Therefore, these structures are excluded from all further investigations. With the remaining 96 pyrazines the best results for classification are obtained with a multilayer perceptron network architecture with five input neurons and one hidden layer containing three neurons. The ANN is able to distinguish between these three groups of aroma impressions by using the following significant descriptors as inputs: sum of electrotopological indices, number of carbon atoms of the substituent R2, charge on the first atom of the substituent R4, and the molecular surface of the substituents R3 and R1. The values of these descriptors are shown in Table 2. In Table 3 the corresponding correlation matrix is depicted.

These descriptors are chosen by using a sensitivity analysis, which gives some information about the relative importance of the variables used for training the neural network. This analysis tests how the predictive ability of the ANN would change if the respective input variables are unavailable. The data set is submitted to the network repeatedly, with each variable in term treated as missing, and the resulting network error is recorded. If an important variable is removed in this procedure, the error will increase significantly.

The number of neurons in the hidden layer is determined by trial and error, taking into account the empirical rule mentioned by So and Richards, based on the ρ value (ρ is equal to the quotient between the number of data points in the training set and the number of adjustable weights controlled by the network).[36] The range $1.8 < \rho < 2.2$ has been suggested as an empirical guideline of acceptable ρ values. It has been defined that for $\rho < 1.0$ the network simply memorizes the data while for $\rho > 3.0$ the network is not able to generalize.

The compounds are split randomly into three classes: 54 training compounds, 21 verification substances (compounds **3**, **4**, **7**, **26**, **28**, **32**, **34**, **38**, **39**, **45**, **54**, **57**, **59**, **65**, **66**, **68**, **74**, **81**, **84**, **87**, and **93**), and 21 test substances (compounds **2**, **9**, **10**, **11**, **23**, **25**, **27**, **36**, **40**, **42**, **43**, **46**, **49**, **52**, **60**, **63**, **67**, **70**, **79**, **95**, and **97**). The output is defined as a nominal variable (green, nutty, or bell-pepper). The training of the network is performed with a learning rate of 0.1, a momentum factor of 0.3, and stopping conditions with a minimum improvement of 0.01 for both the training and the verification error.

After the network is run, a verification error of 0.208, a training error of 0.182, and a test error of 0.304 are observed. The errors are defined as the sum of the squared differences between the predicted and actual output values on each output unit. The performance of the network with a correct classification of 95.2% for the verification set, 96.2% for the training, and 85.7% for the test set is quite impressive. Table 4 demonstrates that only 6 out of 96 pyrazines are classified wrongly. Compounds **27**, **30**, and **32** with a green flavor are predicted as bell-pepper-smelling pyrazines. This may be explained by the different molecular surface values of the substituent R3, which are quite high for the green-smelling compounds but show low values in the case of the three misclassified ones. Furthermore, these pyrazines contain a higher number of carbon atoms at substituent R2, which would again indicate a bell-pepper aroma impression. The two nutty-smelling pyrazines **46** and **55** are classified into the green aroma group. This seems to be due to their electrotopological indices, which are relatively high in comparison to the values of the green aroma substances. Additionally, one green aroma compound is predicted to have a nutty aroma impression (compound **23**), which is caused by low electrotopological indices normally describing nutty compounds.

Table 2. List of Calculated Values of All Significant Descriptors (SEI, NrC2, CR4, MS3, MS1 of the Classification Model, CR1, NrC1, NrO1 of the Green Model, and CR3, MS, NrO1 of the Bell-Pepper Set), Where Unit for Charge = C, Unit for Surface = Å

compd	CR1[a]	CR3[b]	CR4[c]	MS[d]	MS1[e]	MS3[f]	SEI[g]	NrC1[h]	NrC2[i]	NrO1[j]
1	−0.9353	0.2490	−0.4058	211.198	78.969	18.096	26.167	2	0	0
2	−0.7128	0.2642	0.2692	182.346	107.174	18.096	23.667	4	0	1
3	−0.8232	−0.3676	0.2661	245.581	111.735	82.490	35.500	6	1	0
4	0.4853	−0.4444	0.2664	274.595	87.234	139.104	28.967	2	1	0
5	−0.8099	0.2576	0.2583	170.651	80.648	18.096	21.333	2	1	0
6	−0.7254	−0.4001	0.2582	202.892	51.852	99.369	27.333	1	1	1
7	−0.7254	−0.4017	0.2582	222.871	51.852	124.327	28.833	1	1	1
8	−0.7256	−0.4013	0.2581	240.544	51.852	144.298	30.333	1	1	1
9	−0.7292	−0.3448	0.2575	226.526	72.739	99.565	28.833	2	1	1
10	−0.7291	−0.4042	0.2583	226.425	72.688	101.011	28.833	2	1	1
11	−0.7282	−0.4090	0.2591	236.395	73.304	122.060	30.333	2	1	1
12	−0.7281	−0.4478	0.2568	260.072	72.627	144.740	31.833	2	1	1
13	−0.8234	−0.4057	0.2633	264.057	111.945	101.039	37.000	6	1	1
14	−0.8234	−0.4424	0.2621	298.833	112.365	138.510	40.000	6	1	1
15	0.4055	−0.4085	0.2612	220.598	67.841	98.605	24.467	1	1	0
16	0.4048	−0.4050	0.2620	256.242	67.839	142.898	27.467	1	1	0
17	0.4077	−0.4051	0.2620	252.068	87.391	123.708	27.467	2	1	0
18	−0.7481	−0.5913	0.2621	206.592	52.324	43.251	31.750	1	3	1
19	−0.7497	0.2589	−0.5948	207.829	52.531	18.096	31.750	1	3	1
20	−0.7115	0.2618	−0.5990	192.617	52.905	18.096	31.167	1	2	1
21	−0.7373	−0.7286	−0.5813	200.959	52.391	52.226	36.333	1	2	1
22	0.2569	0.2563	−0.5938	161.328	18.096	18.096	18.833	0	2	0
23	−0.4585	0.2585	0.2585	175.352	64.196	18.096	20.333	2	2	0
24	0.2544	−0.6089	−0.6058	198.336	18.096	43.101	22.333	0	3	0
25	0.2605	0.2622	0.2630	180.109	18.096	18.096	20.167	0	4	0
26	0.2613	0.2625	0.2633	174.403	18.096	18.096	20.500	0	4	0
27	0.4063	0.2654	0.2674	204.772	68.009	18.096	22.800	1	4	0
28	0.2661	0.2614	0.2621	193.312	18.096	18.096	21.667	0	5	0
29	0.2538	−0.6092	−0.6058	226.661	18.096	43.156	25.000	0	5	0
30	−0.7254	0.2598	0.2629	211.445	51.956	18.096	26.833	1	5	1
31	−0.6083	−0.5954	0.2527	219.394	43.277	43.361	25.333	1	5	0
32	−0.7254	0.2599	0.2629	252.916	51.852	18.096	29.833	1	7	1
33	−0.6088	−0.6088	0.2576	140.602	43.040	43.309	17.333	1	0	0
34	−0.7053	0.2651	0.2703	133.262	52.309	18.096	19.167	1	0	1
35	−0.7081	0.2584	−0.5950	151.921	52.359	18.096	20.833	1	0	1
36	−0.7086	0.2645	0.2695	148.675	71.745	18.096	20.667	2	0	1
37	0.4785	0.2700	0.2740	143.139	67.778	18.096	16.300	1	0	0
38	0.4738	0.2639	−0.6111	160.160	67.581	18.096	17.967	1	0	0
39	0.4911	0.2691	0.2730	161.515	86.415	18.096	17.800	2	0	0
40	−0.6068	0.2599	0.2598	141.014	42.826	18.096	17.333	1	1	0
41	−0.6058	−0.5947	0.2536	161.421	43.089	42.880	19.000	1	1	0
42	−0.6070	−0.6070	−0.6070	179.594	42.520	42.748	20.667	1	1	0
43	−0.8481	0.2568	0.2586	150.729	59.433	18.096	19.833	1	1	0
44	−0.7247	0.2603	0.2632	152.611	52.018	18.096	20.833	1	1	1
45	−0.7270	0.2537	−0.5941	173.384	52.068	18.096	22.500	1	1	1
46	−0.7282	0.2600	0.2629	169.301	73.457	18.096	22.333	2	1	1
47	0.4119	0.2658	0.2677	157.587	67.221	18.096	17.967	1	1	0
48	0.4249	0.2646	0.2665	176.707	85.564	18.096	19.467	2	1	0
49	0.2607	0.2626	0.2633	142.854	18.096	18.096	17.167	0	2	0
50	0.2540	−0.6090	−0.6058	179.823	18.096	43.153	20.500	0	2	0
51	−0.6112	0.2531	−0.5949	178.865	42.898	18.096	20.500	1	2	0
52	−0.6079	−0.5952	0.2530	178.852	43.105	42.936	20.500	1	2	0
53	0.4102	0.2649	0.2669	195.782	87.651	18.096	20.967	2	2	0
54	0.2546	−0.6091	−0.6058	211.111	18.096	43.208	23.833	0	4	0
55	0.5557	0.2638	0.2678	328.729	126.439	18.096	38.133	6	8	0
56	−0.6093	0.2588	0.2590	171.619	42.740	18.096	20.333	1	3	0
57	−0.7256	0.2597	0.2628	182.867	53.385	18.096	23.833	1	3	1
58	0.4075	0.2646	0.2669	192.633	67.811	18.096	20.967	1	3	0
59	−0.6102	0.2586	0.2591	170.045	42.310	18.096	20.667	1	3	0
60	−0.7233	0.2599	0.2632	179.564	53.485	18.096	24.167	1	3	1
61	−0.7256	0.2534	−0.5939	199.857	53.383	18.096	25.833	1	3	1
62	−0.7237	0.2572	−0.5902	200.018	53.477	43.303	25.833	1	3	1
63	−0.7260	−0.7279	−0.5755	223.599	53.571	53.625	31.000	1	3	1
64	−0.7255	−0.5726	−0.7264	224.199	53.888	43.055	31.000	1	3	1
65	−0.7262	−0.7304	−0.3641	253.784	53.522	53.731	34.333	1	3	1
66	0.4069	0.0377	0.0477	188.758	67.762	18.096	21.300	1	3	0
67	−0.7260	0.2594	0.2627	202.189	53.015	18.096	25.333	1	4	1
68	0.4216	0.2636	0.2658	222.354	87.782	18.096	23.967	2	4	0
69	−0.6118	0.2593	0.2595	188.415	42.574	18.096	22.167	1	4	0
70	−0.7259	0.2598	0.2631	196.822	53.431	18.096	25.667	1	4	1
71	−0.7283	0.2533	−0.5940	217.152	53.430	18.096	27.333	1	4	1
72	−0.7264	−0.5905	0.2570	212.014	53.685	43.409	27.333	1	4	1
73	−0.7288	−0.6041	−0.6061	230.573	53.794	42.784	29.000	1	4	1
74	−0.7234	0.2597	0.2630	196.510	53.325	18.096	25.667	1	4	1
75	−0.7285	0.2589	0.2623	241.198	72.946	18.096	28.333	2	5	1

Table 2 (Continued)

compd	CR1[a]	CR3[b]	CR4[c]	MS[d]	MS1[e]	MS3[f]	SEI[g]	NrC1[h]	NrC2[i]	NrO1[j]
76	0.4075	0.2644	0.2667	218.706	67.811	18.096	23.967	1	5	0
77	0.4102	0.2643	0.2667	236.390	87.391	18.096	25.467	2	5	0
78	−0.7252	0.2605	0.2633	210.145	53.118	18.096	27.167	1	5	1
79	−0.7248	0.2610	0.2642	208.756	54.102	18.096	27.167	1	5	1
80	−0.7260	0.2604	0.2633	217.960	53.385	18.096	28.333	1	5	1
81	−0.7276	0.2603	0.2633	201.299	53.533	18.096	26.333	1	4	1
82	−0.7255	0.2606	0.2636	195.675	53.224	18.096	26.333	1	4	1
83	−0.7260	0.2594	0.2627	238.011	52.960	18.096	28.333	1	6	1
84	−0.7260	0.2594	0.2627	227.525	53.169	18.096	28.667	1	6	1
85	−0.7261	0.2601	0.2633	228.451	53.692	18.096	28.667	1	6	1
86	−0.7260	0.2593	0.2627	274.687	52.853	18.096	31.333	1	8	1
87	−0.7295	0.2590	0.2624	290.243	74.110	18.096	32.833	2	8	1
88	0.4073	0.0323	0.2667	275.203	67.811	18.096	28.467	1	8	0
89	0.4102	0.2643	0.2667	292.385	87.392	18.096	29.967	2	8	0
90	−0.7260	0.2593	0.2626	312.371	53.068	18.096	34.333	1	10	1
91	−0.7295	0.2590	0.2624	327.420	72.899	18.096	35.833	2	10	1
92	−0.7277	0.7277	−0.5757	196.349	53.580	53.625	27.667	1	1	1
93	−0.7252	0.2598	0.2628	165.935	53.333	18.096	22.333	1	2	1
94	−0.7266	0.2593	0.2626	215.834	53.644	18.096	27.167	1	5	1
95	−0.7257	0.2594	0.2627	280.239	53.333	18.096	33.167	1	9	1
96	−0.7254	0.2602	0.2635	286.815	53.483	18.096	33.167	1	9	1
97	−0.7284	0.2542	−0.4036	274.208	53.375	18.096	32.167	1	4	1
98	−0.7284	0.2593	0.2626	215.980	73.889	18.096	27.167	2	4	1

[a] Charge of the first atom of substituent R1. [b] Charge of the first atom of substituent R3. [c] Charge of the first atom of substituent R4. [d] Molecular surface of the whole molecule. [e] Molecular surface of substituent R1. [f] Molecular surface of substituent R3. [g] Sum of electrotopological indices. [h] Number of carbon atoms of substituent R1. [i] Number of carbon atoms of substituent R2. [j] Number of oxygen atoms of substituent R1.

Table 3. Correlation Matrix of the Significant Descriptors of the Classification Model

	CR4	MS1	MS3	SEI	NrC2
CR4	1				
MS1	0.285	1			
MS3	0.064	0.246	1		
SEI	−0.055	0.365	0.370	1	
NrC2	0.167	−0.113	−0.385	0.452	1

Table 4. Classification of the Training, Verification, and Test Sets of Pyrazines of Bell-Pepper, Nutty, and Green Fragrances

	training			verification			test		
	bp	nutty	green	bp	nutty	green	bp	nutty	green
total	26	11	17	10	5	6	7	7	7
correct	26	10	16	10	5	5	7	6	5
wrong	0	1	1	0	0	1	0	1	2
unknown	0	0	0	0	0	0	0	0	0

Table 5. Test Set of Green-Smelling Pyrazines

structure	predicted log(1/*c*)	actual log(1/*c*)	residual
6	6.800	6.000	0.800
9	4.406	4.209	0.196
14	2.709	3.285	−0.576
28	3.521	4.477	−0.956
32	6.800	6.903	−0.103

2. Prediction of Threshold Values of Pyrazines with Green Flavor. A more detailed consideration of pyrazine-derived aroma compounds with green odor impressions can be performed quantitatively using their threshold values. The aim of this particular study is to find a structural principle of this aroma impression and to estimate those parameters that are of importance for this specific flavor. A total of 24 pyrazine derivatives with this aroma are selected because their biological activities (log(1/*c*) values) also have been estimated quantitatively. The randomly selected verification set consists of compounds **4**, **7**, **12**, **15**, and **17**, and the test set consists of structures **6**, **9**, **14**, **28**, and **32** (see Table 1). A multilayer perceptron network architecture is applied, and the back-propagation algorithm is used for training. The best model is received with one hidden layer containing three neurons by using the following three inputs (in order of decreasing importance) that were obtained from sensitivity analysis: charge on the first atom of substituent R1, number of carbon atoms of substituent R1, and number of oxygen atoms of substituent R1. With these significant parameters a network with a training error of 0.758, a verification error of 0.533, and a test error of 0.622 are obtained. Furthermore, we obtain a Pearson *R* correlation coefficient of 0.859 for the training set, 0.918 for the verification, and 0.948 for the test set (Table 6). By using eq 1 for the calculation of the predictive power Q^2, we yield a Q^2 of 0.771 for the test set.

$$Q^2 = 1 - \frac{\sum(\text{actual} - \text{predicted})^2}{\sum(\text{actual} - \text{mean})^2} \qquad (1)$$

Figure 2 shows a plot of the experimental log(1/*c*) values of the training and the verification sets against the calculated log(1/*c*) values. In Table 5 the actual and the predicted log(1/*c*) values of the test set are presented, where four out of the five structures are quite well predicted.

3. Prediction of Threshold Values of Pyrazines with Bell-Pepper Aroma Impression. For comparison, bell-pepper aroma compounds are considered in the same way. To predict the odor threshold values of 37 bell-pepper aroma compounds, the data set is split into the training set, the verification set (containing substances **57**, **60**, **67**, **70**, **72**, **80**, **82**, **90**, **92**), and the test set (substances **56**, **61**, **63**, **71**, **77**, **83**, **85**, **87**, **88**). Best results are obtained using a multilayer perceptron network architecture with two hidden layers containing two and one neurons, respectively. When the training and the verification sets are run with a back-propagation algorithm, an excellent Q^2 value (see eq 1) of 0.800

Table 6. Statistical Parameters of Pyrazines with Bell-Pepper and Green Aroma

	bell-pepper			green		
	training log(1/*c*)	veritifcation log(1/*c*)	test log(1/*c*)	training log(1/*c*)	verification log(1/*c*)	test log(1/*c*)
data mean[a]	5.314	5.644	5.441	3.888	4.971	4.975
data SD[b]	1.698	1.506	1.355	1.507	1.406	1.454
error mean[c]	0.001	0.005	0.188	0.118	−0.092	−0.127
error SD[d]	0.596	0.631	0.565	0.777	0.587	0.680
abs error mean[e]	0.477	0.484	0.480	0.543	0.460	0.526
SD ratio[f]	0.351	0.418	0.417	0.516	0.417	0.468
correlation[g]	0.936	0.912	0.926	0.859	0.918	0.948

[a] Average value of the target output variable. [b] Standard deviation of the target output variable. [c] Average error (residual) between target and actual output values of the output variable. [d] Average absolut error (difference between target and actual output values) of the output variable. [e] Standard deviation of errors for the output variable. [f] Error data standard deviation ratio. [g] Standard Pearson *R* correlation coefficient between the target and actual output values.

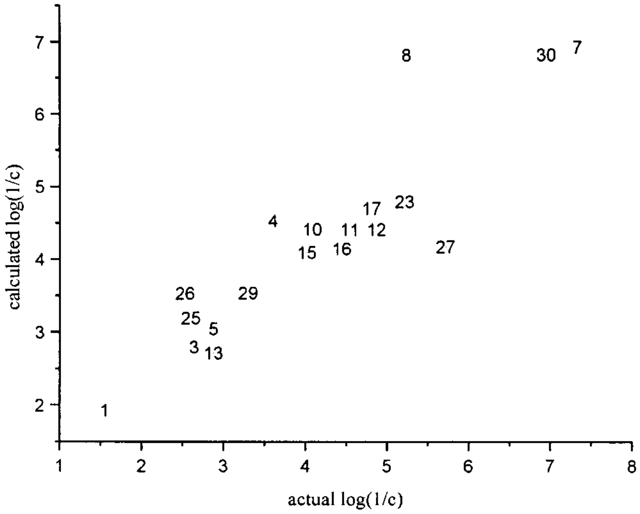


Figure 2. Experimental log(1/*c*) values plotted against predicted log(1/*c*) values of green-smelling pyrazines obtained from neural network containing training and verification sets.

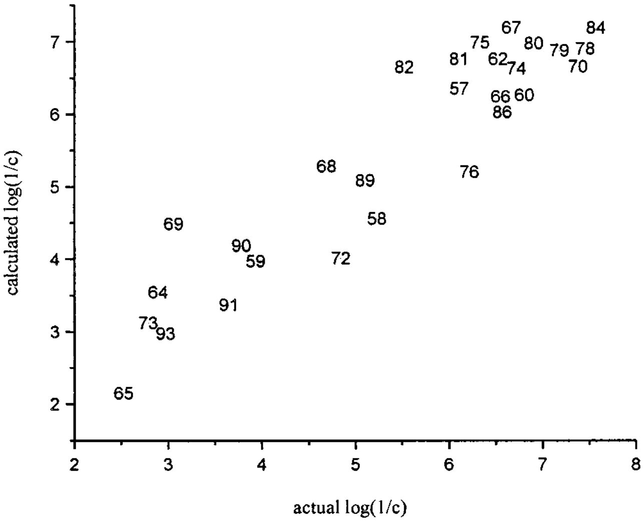


Figure 3. Experimental and calculated log(1/*c*) values of training and verification sets of pyrazines with bell-pepper flavor.

Table 7. Experimental and Predicted log(1/*c*) Values of the Bell-Pepper Test Set by Neural Network

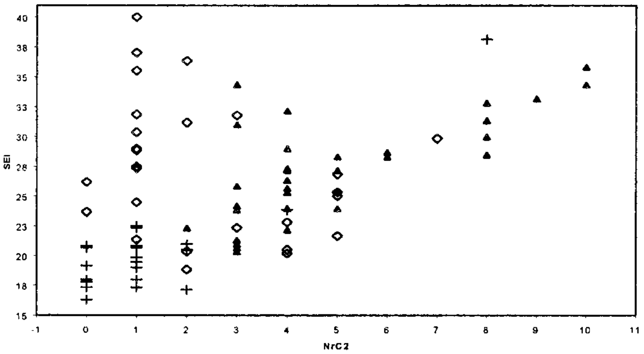
structure	predicted log(1/*c*)	actual log(1/*c*)	residual
56	4.013	3.356	0.657
61	6.787	6.567	0.219
63	2.804	3.477	−0.643
71	7.014	5.840	1.173
77	5.509	5.322	0.186
83	6.980	6.443	0.537
85	7.023	7.385	−0.362
87	5.305	5.072	0.232
88	5.225	5.532	−0.306

for the test set and rather high Pearson *R* correlation coefficients (for the training set $R = 0.936$, for the verification set $R = 0.912$, and for the test set $R = 0.926$) are obtained (Table 6, Figure 3).

The following three molecular properties were classified as important, arranged in decreasing influence: the charge of the first atom of substituent R3, the molecular surface, and the number of oxygen atoms of substituent R1. We receive an error for the training set of 0.581, for the verification set of 0.594, and finally for the test set of 0.565. In Table 7 the actual and predicted log(1/*c*) values of nine test molecules are presented. It can be observed that eight out of nine compounds are quite well predicted.

Discussion

The application of ANNs on a series of aroma compounds enables us to distinguish among the three aroma impressions (green, bell-pepper, and nutty) by using only five descriptors. The aroma impression is preferentially determined by the sum of the electrotopological indices, which gives information related to the electronic and topological state of the atoms in the molecule.[37] A low value for the sum of the electrotopological indices (~15−21) indicates a nutty aroma impression, while a higher one (~22−30) leads to a green aroma. The bell-pepper aroma impression cannot be classified by using this descriptor. The distinction between the aroma impressions of green, bell-pepper, and nutty is additionally influenced by the molecular surface of the substituents R1 and R3 and, less important, by the number of carbon atoms of substituent R2 and the charge of the first atom of substituent R4. Green-smelling pyrazines generally show a higher value for the molecular surface of substituent R3 than the other two impressions. On the other hand, the values of the molecular surface of substituent R1 of the bell-pepper group are within the range 45−55 Å^2, while the other two classes show widespread values for this descriptor. The number of carbon atoms of substituent R2 characterizes the group of bell-pepper pyrazines. If the side chain at this position contains more than two carbon atoms (3−10) the substance has a strong tendency to have bell-pepper aroma, whereas nutty-smelling aroma compounds have only a hydrogen atom, a methyl group, or an ethyl group

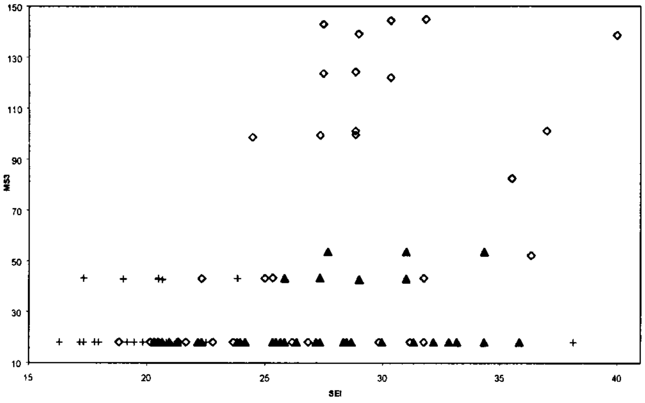


+ nutty, ▲ bellpepper, ◊ green

NrC2 = number of carbon atoms of substituent R2,

SEI = sum of electrotopological indices

Figure 4. Clustering of 96 aroma compounds by using the number of carbon atoms of substituent R2 and the sum of electrotopological indices as descriptors.



+ nutty, ▲ bellpepper, ◊ green

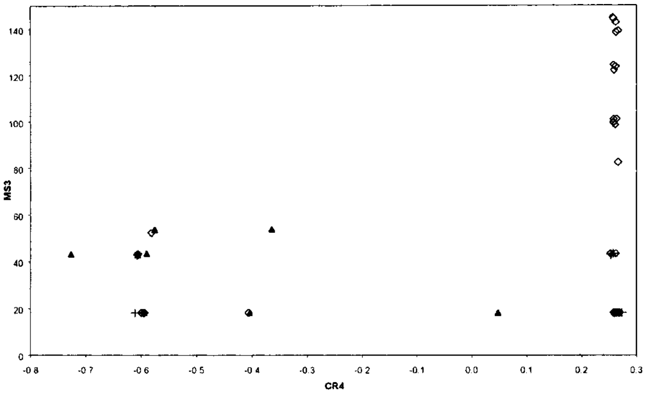
SEI = sum of electrotopological indices

MS3 = molecular surface of substituent R3

Figure 5. Clustering of 96 aroma compounds by using the sum of electrotopological indices and the molecular surface of substituent R3 as descriptors.

as the substituent at position R2. In Figures 4–6 the odor quality domains are depicted as functions of two physicochemical properties. It can be shown that a clustering with only two significant descriptors is not possible, which clearly indicates a nonlinear dependency. For the threshold prediction other descriptors are required. The aroma intensity of green-smelling pyrazines is highly influenced by the charge of the first atom of substituent R1. The more negative the charge is at this position, the higher is the odor intensity of these pyrazines. Furthermore, the presence of a methoxy group at the position of R1 results in a lower threshold.

The prediction of the threshold values of pyrazines with bell-pepper impression is predominantly influenced by the charge of the first atom of substituent R3. A higher biological activity can be obtained with a charge in the range −0.3 to +0.3 C. Another important descriptor is the molecular surface of the whole molecule. The lowest threshold values are observed for molecules with a molecular surface between 200 and 250 Å^2. These parameters were also found to be important by QSAR and CoMFA investigations.[15] Additionally, an oxygen atom at substituent R1 favors the bell-pepper-smelling impression of pyrazines.



+ nutty, ▲ bellpepper, ◊ green

CR4 = charge of the first atom of substituent R4,

MS3 = molecular surface of substituent R3

Figure 6. Clustering of 96 aroma compounds by using the charge of the first atom of substituent R4 and the molecular surface of substituent R3 as descriptors.

From the examples given above it can be concluded that the treatment of the structure–flavor relationships of aroma compounds by ANNs leads to rather reliable prediction models. Both the classification model and the prediction model show good agreement between the experimental properties and the calculated information. In particular the classification models can only be treated by neural networks as a consequence of nonlinearities in the influence of some descriptors.

References

(1) Flower, D. H. The lipocalin protein family: structure and function. *Biochem. J.* **1996**, *318*, 1–14.

(2) Bignetti, E.; Cavaggioni, A.; Pelosi, P.; Persaud, K. C.; Sorbi, R. T.; Tirindelli, R. Purification and characterisation of an odorant binding protein. *Eur. J. Biochem.* **1985**, *149*, 227–231.

(3) Pevsner, J.; Hwang, P. M.; Sklar, P. B.; Venable, J. C.; Snyder, S. H. Odor binding protein and its mRNA are localized to lateral nasal gland implying a carrier function. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2383–2387.

(4) Pes, D.; Dal Monte, M.; Ganni, M.; Pelosi, P. Isolation of two odorant binding proteins from mouse nasal tissue. *Comp. Biochem. Physiol.* **1992**, *103B*, 1011–1017.

(5) Paolini, S.; Scaloni, A.; Amoresano, A.; Marchese, S.; Napolitano, E.; Pelosi, P. Amino acid sequence, post translational modifications, binding and labeling of porcine odorant binding protein. *Chem. Senses* **1998**, *23*, 689–698.

(6) Tegoni, M.; Pelosi, P.; Vincet, F.; Spinelli, S.; Campanacci, V.; Grolli, S.; Cambillau, C. Mammalian odorant binding proteins. *Biochim. Biophys. Acta* **2000**, *1482*, 229–240.

(7) Kubinyi, H.; Kehrhahn, O. H. Quantitative structure–activity relationships. VI. Nonlinear dependence of biological activity on hydrophobic character: calculation procedures for bilinear model. *Arzneim.-Forsch.* **1978**, *28*, 598–601.

(8) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, New York, 1999.

(9) Chastrette, M.; de Saint Laumer, J. Y. Adapting the structure of a neural network to extract chemical information. Application to structure odor relationships. *SAR QSAR Environ. Res.* **1992**, *1*, 221–231.

(10) Chastrette, M.; Zakarya, D.; Peyraud, J. F. Structure musk odour relationship studies for tetralins and indans using neural networks. *Eur. J. Med. Chem.* **1994**, *29*, 343–348.

(11) Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, J. M.; Chastrette, M.; Zakarya, D. Structure musk odour relationship studies of tetralins and indan compounds using neural networks. *New. J. Chem.* **1998**, *22*, 839–843.

(12) Zakarya, D.; Chastrette, M.; Tollabi, M.; Fkih-Tetouani, S. Structure–camphour odour relationships using the Generation and Selection of Pertinent Descriptors Approach. *Chemom. Intell. Lab. Syst.* **1999**, *48*, 35–46.

(13) Zakarya, D.; Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, J. M. Application of neural networks to structure sandalwood odour relationships. *J. Phys. Org. Chem*. **1997**, *10*, 612–622.

(14) Chastrette, M.; El Aidi, C. Structure–Bell-Pepper Odour Relationships for Pyrazines and Pyridines Using Neural Networks. In *Neural Networks in QSAR and Drug Design*; Devillers, J., Ed.; Academic Press: London, 1996; pp 83–96.

(15) Klein, Ch. Th.; Wailzer, B.; Buchbauer, G.; Wolschann, P. Threshold-Based Structure–Activity Relationships of Pyrazines with Bell-Pepper Flavor. *J. Agric. Food Chem*. **2000**, *48*, 4273–4278.

(16) Masuda, H.; Mihara, S. Olfactive Properties of Alkylpyrazines and 3-Substituted 2-Alkylpyrazines. *J. Agric. Food Chem*. **1988**, *36*, 584–587.

(17) Buttery, R. G.; Guadagni, G. D.; Ling, L. C. Volatile Components of Baked Potatoes. *J. Sci. Food Agric.* **1973**, *24*, 1125–1131.

(18) Parliament, T. H.; Epstein, M. F. Organoleptic Properties of Some Alkyl-Substituted Alkoxy- and Alkylthiopyrazines. *J. Agric. Food Chem*. **1973**, *21*, 714–716.

(19) *Hyperchem*, version 5.0; Hypercube Inc.: Gainesville, FL, 1997.

(20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. *Gaussian 94*, revision B.3; Gaussian, Inc.: Pittsburgh, PA, 1995.

(21) *TSAR*, version 3.2; Oxford Molecular, Ltd.: Oxford, England, 1999.

(22) *Trajan Neural Networks*, version 4.0; Trajan Software, Ltd.: Durham, U.K., 1999.

(23) Cerny, C.; Grosch, W. Z. Quantification of character-impact odour compounds of roasted beef. *Z. Lebensm.-Unters. Forsch.* **1993**, *196*, 417–422.

(24) Masuda, H.; Mihara, S. Synthesis of Alkoxy-, (Alkylthio)-, Phenoxy-, and (Phenylthio)pyrazines and Their Olfactive Properties. *J. Agric. Food Chem*. **1986**, *34*, 377–381.

(25) Shibamoto, T. Odor Threshold of Some Pyrazines. *J. Food Sci.* **1986**, *51*, 1098–1099.

(26) Seifert, R. M.; Buttery, R. G.; Guadagni, D. G.; Black, D. R.; Harris; J. G. Synthesis of Some 2-Methoxy-3-alkylpyrazines with Strong Bell-Pepper-like Odors. *J. Agric. Food Chem*. **1970**, *18*, 246–249.

(27) Takken, H. J.; Van der Linde, M. L.; Boelens, M.; Van Dort, J. M. Olfactive Properties of a Number of Polysubstituted Pyrazines. *J. Agric. Food Chem*. **1975**, *23*, 638–642.

(28) Boelens, M. H.; Van Gemert, L. J. Structure–Activity Relationships of Natural Volatile Nitrogen Compounds. *Perfum. Flavor.* **1995**, *20*, 63–76.

(29) Pittet, A. O.; Hruza, D. E. Comparative Study of Flavor Properties of Thiazole Derivatives. *J. Agric. Food Chem*. **1974**, *22*, 264–269.

(30) Calabretta, P. J. Synthesis of Some Substituted Pyrazines and Their Olfactory Properties. *Perfum. Flavor.* **1978**, *3* (3), 33–42.

(31) Fors, S. M.; Olofsson, B. K. Alkylpyrazines, volatiles formed in the Maillard reaction. 1. Determination of odor detection threshold and odor intensity function by dynamic olfactometry. *Chem. Senses* **1985**, *10*, 287–296.

(32) Mihara, S.; Masuda, H.; Tateba, H.; Tuda, T. Olfactive Properties of 3-Substituted 5-Alkyl-2-methylpyrazines. *J. Agric. Food Chem*. **1991**, *39*, 1262–1264.

(33) Mihara, S.; Masuda, H. Structure–Odor Relationships for Disubstituted Pyrazines. *J. Agric. Food Chem*. **1988**, *36*, 1242–1247.

(34) Flament, I.; Stoll, M. Pyrazines. 1. Synthesis of 3-Alkyl-2-methylpyrazines by Condensation of Ethylenediamine with 2,3-Dioxoalkanes. *Helv. Chim. Acta* **1967**, *50*, 1754–1758.

(35) Murray, K. E.; Whitfield, F. The Occurrence of 3-Alkyl-2-methoxypyrazines in Raw Vegetable. *J. Sci. Food Agric.* **1975**, *26*, 937–986.

(36) So, S.; Richards, W. G. Application of neural networks: Quantitative structure–activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors. *J. Med. Chem*. **1992**, *35*, 3201–3207.

(37) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: London, 1999.

JM001129M