Bayesian Neural Networks for Aroma Classification

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

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

Received March 13, 2002

Bayesian Neural Networks (BNNs) are investigated to test their potential to distinguish between different aroma impressions. Special attention is thereby drawn on mixed aroma impressions, resulting from the flavor description of a single compound with more than one aroma quality. The structures of 133 pyrazine-derived aroma compounds as well as their aroma descriptions are selected for comparison. The information fed into the neural networks is based on molecular descriptors calculated from the geometrically optimized chemical structures. While in the case of the Probabilistic Neural Network (PNN) the networks' output consists of a categorical variable, the output for the General Regression Neural Network (GRNN) is defined in a numerical way. The best models attain comparable performance with a correct prediction of 90.8% of the cases for PNN and 89.9% for GRNN, respectively. Comparison of the BNN results to those obtained by Multiple Linear Regression (MLR) points out that the nonlinear methods work significantly better on the studied problem and that BNNs can be applied to multiple-category problems in structure-flavor relationships with good accuracy.

INTRODUCTION

Aroma compounds are volatile, hydrophobic compounds. It has been proven that they bind to specific receptors localized in the olfactory mucosa. In mammals, these olfactory receptors (ORs) are immersed in the protective nasal mucus and thus are not in direct contact with air. As aroma compounds are hydrophobic molecules they cannot easily cross the hydrophilic barrier of the nasal mucus. Therefore, so-called odorant binding proteins (OBPs) are necessary for the transportation of the aroma compounds to the ORs through the aqueous barrier of the mucus.[1] OBPs have been identified in the olfactory mucosa of different animals.[2−5] The three-dimensional structures of bovine and porcine OBP have been determined by X-ray investigations.[6] While OBPs are identified as members of the lipocalin superfamily, the ORs pertain to a large multigene family containing about 1000 members. They belong to the class of the seven-helix transmembrane G-protein coupled receptors and can be divided into several subfamilies.[7] In 1999, Malnic et al.[8] provided evidence that the mammalian olfactory system uses a combinatorial receptor coding scheme to encode odor identities and to discriminate various odor impressions. They showed that single receptors can recognize multiple odorants. Furthermore, a single odorant is typically recognized by multiple receptors, and different odorants are recognized by different combinations of receptors. Therefore, the capacity of the mammalian olfactory system to distinguish between odor qualities is very high.

Nevertheless, the gap between the knowledge of the primary structure and the three-dimensional geometry of ORs is large: while the sequences of some receptors are known, no detailed structural elucidation exists up to now. Therefore, the direct study of ligand−receptor interactions in the case of olfaction is not feasible, but quantitative structure−activity relationship (QSAR) approaches can be applied, as they deal with this situation indirectly. They correlate the biological activity of the ligands with their structural or physicochemical properties and extend the correlated properties for the prediction of new active ligands.[9] QSAR methods have been successfully applied to many drug design problems and should also be helpful for the study of interactions between aroma compounds and the odorant binding protein as well as the olfactory receptors. "Traditional" QSAR approaches are based on linear regression methods such as multiple linear regression (MLR) and partial least squares (PLS). Sometimes, such linear methods lead to unsatisfactory results due to nonlinear relationships within the studied data set. In such cases, regression methods based on artificial neural networks (ANNs) can be applied, which are able to construct nonlinear decision boundaries. Thus, they are well suited for the categorization of the groups of patterns which do not present Gaussian shapes in the feature space.[9]

The aim of the presented study is to extend our previous investigation on classification problems of aroma impressions[10] one step further by testing the potency of Bayesian Neural Networks (BNNs) on a set of pyrazines with many categories of different aroma qualities. Pyrazines are nitrogen-containing heterocycles which have been recognized to be important flavor ingredients. They are of interest for the aroma chemistry because of their high odor potency and their characteristic sensory properties. Pyrazines show a broad spectrum of aroma impressions, whereby small differences in the structure of the four substituents and changes in their relative position at the heteroaromatic ring cause significant

* Corresponding author phone: +43 1 4277 52772; fax: 0043 1 4277 9527; e-mail: Karl.Peter.Wolschann@univie.ac.at.
[†] Institute of Theoretical Chemistry and Structural Biology.
[‡] Institute of Pharmaceutical Chemistry.

10.1021/ci0202640 CCC: $22.00
Published on Web 10/10/2002

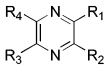

Figure 1. General structure of pyrazines.

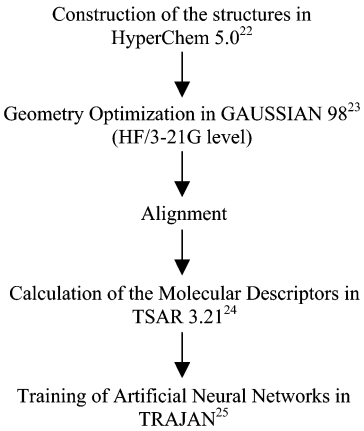

Figure 2. Flowchart of the initial structure treatment.

modifications of the aroma properties. In the presented study the aroma qualities green, bellpepper, nutty, earthy, and sweet are investigated. Special attention is drawn on aroma mixtures of two or more different aroma impressions. Real mixtures of such varying aroma qualities are taken into consideration as well as dominant aroma impressions with weaker tonalities. It can be shown that—in comparison to Multiple Linear Regression (MLR)—both types of BNNs prove to be useful classification and prediction tools for the discrimination between the different sensory properties.

MATERIALS AND METHODS

The general structure of pyrazines can be seen from Figure 1. The molecular structures and olfactory properties of the 133 pyrazines in our sample are taken from literature[11−21] and are available as Supporting Information. From the many different notes which are used to describe the aroma impressions, the qualities green, bellpepper, nutty, earthy, and sweet are investigated. As already mentioned, many authors describe the aroma property of a molecule with more than one note (i.e. green and nutty). Therefore, also aroma mixtures are taken into considerations (i.e. green-nutty). The initial treatment of the selected structures is reported in Figure 2. Before calculating the physicochemical parameters, the structures are superimposed by taking the pyrazine ring as common backbone. By convention we decide that all substituents containing a heteroatom should occupy position R_1. For structures without a substituent with heteroatom, the methyl group is assumed to be located at position R_1. If both of these structural features are missing, the longest side-chain is laid on position R_2. With these decision rules the alignment of all compounds is exactly defined.

Molecular Descriptors. The steric properties of the substituents are described i.e., by molecular mass, molecular surface, molecular volume, and Verloop parameters. The number of carbon atoms of every substitutent and the number of oxygen and sulfur atoms at position R_1 are counted. Furthermore, the molecular connectivity Chi and the Kappa shape indices as well as the logP are calculated for all molecules. The electronic effects of the substances are characterized by the Hartree−Fock-derived dipole moments and the point charges on the atoms of the heterocycle as well as on the first atoms of the four substituents R_1-R_4. In sum, a total of 75 molecular descriptors is calculated. This high number of molecular descriptors is reduced during the training process of the neural network by conducting a sensitivity analysis on the networks' input. This analysis is applied on the training set and gives some information about the relative importance of the input variables used. It is tested how the neural network would cope if each of its input variables were unavailable. Therefore, the data set is submitted to the network repeatedly, with each variable in turn treated as missing, and the resulting network error is reported. If an important variable is deleted in this fashion, the error will increase significantly; if an unimportant variable is removed, the error is slightly influenced only. Sensitivity analysis can, therefore, give important insight into the usefulness of individual variables. If a number of models is studied, it is often possible to identify key variables that are always of high sensitivity and others that are commonly of low sensitivity. Consequently, combination of such important variables as input of a final neural network leads to a good model. The notations of the molecular descriptors identified as key variables for the studied networks are presented in Table 1.

Table 1. Notations of the Input Features for PNN and GRNN

input feature	notation
charge of the first atom of the substitutent R_1	R1
lipophilicity of the molecule	logP
dipole moment of the structure	D
molecular surface of the substituent R_4	MS4
no. of oxygen/sulfur atoms of the substituent R_1	#O1/#S1
no. of carbon atoms of the substituent R_2/substituent R_3	#C2/#C3
kappa shape index (third order)	K3

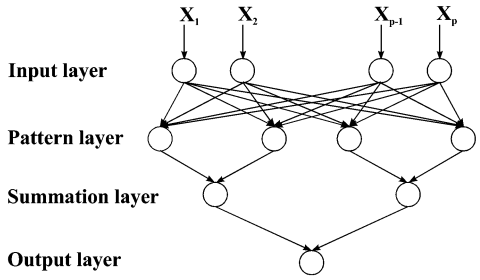

Figure 3. General architecture of Bayesian neural networks.

Calculation Methods. Both types of Bayesian Neural Networks (BNNs) are trained with the TRAJAN[25] software package. Probabilistic Neural Networks (PNNs) and Generalized Regression Neural Networks (GRNNs) were introduced by Specht in 1990 and 1991, respectively.[26,27] While PNNs are generally used for classification problems and, therefore, distinguish between different categories of patterns, GRNNs estimate the most probable value for continuous dependent values. Both network types compute the probability density functions of the given patterns and finally attribute them to the class or value to which they most likely belong.[27] BNNs are feed-forward networks which do not use back-propagation. Their architecture is comprised of four layers as illustrated in Figure 3. The input layer is constituted by a varying number of neurons, which is equal to the number of independent features the network is trained on. The normalized input vector is copied onto the pattern units in the pattern layer, each representing a training case. Instead of the sigmoid activation function commonly applied for back-propagation, BNNs use exponential ones. The activation

Table 2. Number of Cases (Compounds) Assigned to the Various Aroma Impressions and Their output Encoding within the Different Models[a]

PNN/ MLR 1	categorical code	MLR I code	GRNN/ MLR 2	MLR II code	numerical code				
					g	b	n	e	s
22	g	1.0	20	1.0	1	0	0	0	0
34	b	2.0	34	2.0	0	1	0	0	0
0	n	X	3	3.0	0	0	1	0	0
15	e	3.0	15	4.0	0	0	0	1	0
0	s	X	6	5.0	0	0	0	0	1
13	gn	4.0	11	6.0	1	0	1	0	0
15	ge	5.0	14	7.0	1	0	0	1	0
0	gs	X	2	8.0	1	0	0	0	1
3	bn	6.0	3	9.0	0	1	1	0	0
6	be	7.0	5	10.0	0	1	0	1	0
0	ne	X	3	11.0	0	0	1	1	0
12	ns	8.0	8	12.0	0	0	1	0	1
0	es	X	1	13.0	0	0	0	1	1
0	gne	X	2	14.0	1	0	1	1	0
0	gns	X	3	15.0	1	0	1	0	1
0	bns	X	1	16.0	0	1	1	0	1
0	bes	X	1	17.0	0	1	0	1	1
0	nes	X	1	18.0	0	0	1	1	1

[a] g = green, b = bellpepper, n = nutty, e = earthy, s = sweet; 0 = not possessing this aroma quality, 1 = possessing this aroma property.

level from the pattern units is forwarded to the summation unit, where the density estimate on each pattern of each group or possible value is summarized. Finally, a decision with Bayesian theory is established in the fourth layer (decision layer).[26,28,29]

One of the most important factors determining the performance of a BNN is the smoothing constant σ. This parameter controls the size of the receptive region, the field over which the output has a significant response to the input.[30] The smoothing factor should be carefully selected so as to minimize the misclassification rate or the error of the final network. A principle advantage of the BNNs is that they involve a one-pass learning algorithm and are consequently much faster to train than the well-known backpropagation paradigm.[26,27] Furthermore, BNNs differ from classical neural networks in that every weight is replaced by a distribution of weights. This leads to the exploration of a large number of combinations of weights and is less likely to end in a local minimum.[31] Therefore, no test and verification sets are necessary, and in principle all available data can be used for the training of the network. However, it is general practice to hold out some patterns as a test set and to evaluate the performance of the final network through prediction of these compounds.

Moreover, to get some idea about nonlinearity within the studied data set, Multiple Linear Regression (MLR), which establishes a linear combination between the molecular descriptors and the aroma impression of the studied compounds, is applied on the data set.

RESULTS

Table 2 shows the studied aroma impressions as well as the number of molecules assigned to each of these classes. As a consequence of the different working principles of PNNs and GRNNs, the size of the data sets and the assignment of the structures to the various aroma qualities are not completely comparable for the two network types. In the case of PNN, the output is categorical. A structure can only be a member of one of these aroma categories, i.e., it cannot be both green and green-nutty, just green, or green-nutty. When considering mixtures containing more than two sensory properties, the number of cases belonging to a particular high-order class would be too low, and, therefore, no correct prediction of this quality would be possible. For this reason, only two-category mixtures are regarded for training of the PNN, resulting in a data set of 120 pyrazines. In contrast, the main advantage of GRNN is its ability to classify a sample on several classes simultaneously. The output of the GRNN is defined in a numerical way with five dependent variables, one for each of the selected aroma impressions (green, bellpepper, nutty, earthy, and sweet). Training of the GRNN gives information about all five aroma qualities, as it has to decide for every molecule if each of the five aroma impressions is present or absent. It is possible that a structure may get a "presence" for nutty, earthy, and also for sweet, while for the other two properties "absence" is obtained. This can be interpreted as a nutty-earthy-sweet odor, selfevidently without a weight of each aroma impression. Therefore, for GRNN it is also possible to take many-category mixtures into account and to include all 133 pyrazines in the data set.

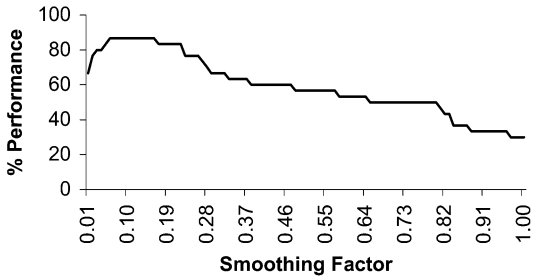

Figure 4. Percentage of correctly classified samples in relationship to the smoothing factor σ.

Probabilistic Neural Network (PNN). Eight different aroma qualities are investigated by PNN (see Table 2), each of them represented by a distinct output unit. The predictive ability of the network is proven by removing 30 compounds from the complete data set, which are used as test set. For training, the smoothing parameter σ is adjusted until an acceptable correct classification rate is obtained, whereby smoothing factors between 0.01 and 1.0 are defined. As presented in Figure 4, small variations within the smoothing constant σ do not change the classification accuracy drastically. Peak accuracy is obtained with every value for σ between 0.06 and 0.16. The final training of the network is performed with a smoothing factor σ of 0.12.

The input of the PNN is obtained by application of sensitivity analysis and consists of six independent features: number of carbon atoms of the substituent R_2, logP, the number of oxygen atoms and the charge of the first atom of the substituent R_1, molecular surface of the substituent R_4, and the Kappa shape index of third order. The contribution of each descriptor to the classification is estimated in Figure 5, which illustrates the variabilities in classification accuracy by removing each of the six input features. The average absolute error for the prediction of the training set by the PNN is 0.149 without removing any feature. This error increases to the highest extent if the number of carbon atoms of the substituent R_2 is canceled from the list of molecular characteristics, which points out that this feature is the most important input variable for this network.

By combination of these six descriptors the PNN reaches a correct classification rate of 92.2% and 86.7% for the

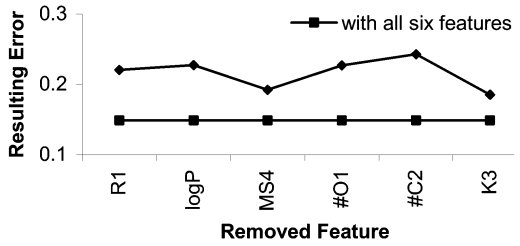

Figure 5. Prediction accuracy of the PNN for the training set when individual features are excluded. The notations of the descriptors are defined in Table 1.

Table 3. Classification Performance of the Training Set for the Different Aroma Impressions by PNN

aroma impression	total	correct	wrong
green	17	15	2
bellpepper	26	25	1
earthy	11	11	0
green-nutty	10	10	0
green-earthy	11	7	4
bellpepper-nutty	2	2	0
bellpepper-earthy	4	4	0
nutty-sweet	9	9	0

Table 4. Prediction of the Aroma Impression of the Test Compounds by PNN and GRNN[a]

			GRNN									
	PNN		green		bellpepper		nutty		earthy		sweet	
compd	act	pre	act	pre	act	pre	act	pre	act	pre	act	pre
1	gn	gn	1	0.45	0	0.00	1	0.50	0	0.65	0	0.03
2	e	gn	0	0.48	0	0.03	0	0.34	1	0.64	0	0.05
3	e	e	0	0.37	0	0.04	0	0.23	1	0.73	0	0.08
4	e	e	0	0.27	0	0.06	0	0.12	1	0.79	0	0.07
5	e	e	0	0.07	0	0.00	0	0.24	1	0.71	0	0.10
6	ge	ge	1	0.94	0	0.00	0	0.42	1	0.27	0	0.08
7	b	b	0	0.03	1	0.95	0	0.01	0	0.05	0	0.00
8	b	b	0	0.04	1	0.96	0	0.00	0	0.08	0	0.00
9	be	be	0	0.02	1	0.98	0	0.02	1	0.09	0	0.00
10	ns	ns	1	0.23	0	0.06	1	0.85	0	0.21	1	0.70
11	g	g	1	0.88	0	0.00	0	0.00	0	0.00	0	0.12
12	g	g	1	0.88	0	0.00	0	0.00	0	0.00	1	0.12
13	gn	g	1	0.99	0	0.00	1	0.00	0	0.00	0	0.00
14	g	g	1	0.99	0	0.00	0	0.00	0	0.00	0	0.00
15	b	b	0	0.12	1	0.87	0	0.01	0	0.07	0	0.00
16	b	b	0	0.36	1	0.63	0	0.25	0	0.25	0	0.09
17	bn	be	0	0.46	1	0.52	1	0.20	0	0.33	0	0.04
18	be	be	0	0.00	1	0.99	0	0.01	1	0.97	0	0.00
19	g	g	1	0.99	0	0.00	0	0.00	0	0.00	1	0.00
20	ns	ns	0	0.02	0	0.34	1	0.84	0	0.17	1	0.56
21	ns	ns	0	0.00	0	0.34	1	0.84	0	0.17	1	0.54
22	ge	ge	1	0.72	0	0.21	0	0.10	1	0.41	0	0.06
23	ge	ge	1	0.83	0	0.00	0	0.17	1	0.93	0	0.00
24	b	b	0	0.09	1	0.91	0	0.00	0	0.24	0	0.00
25	gn	gn	1	0.84	0	0.00	1	0.46	0	0.19	0	0.34
26	g	g	1	0.01	0	0.37	0	0.75	0	0.18	0	0.43
27	ge	g	1	0.72	0	0.20	1	0.11	1	0.42	0	0.08
28	b	b	0	0.04	1	0.95	0	0.00	0	0.07	0	0.00
29	b	b	0	0.02	1	0.97	0	0.00	0	0.07	0	0.00
30	b	b	0	0.03	1	0.94	0	0.03	0	0.07	0	0.00

[a] act = aroma impression given by literature, pre = predicted aroma quality; g = green, b = bellpepper, n = nutty, e = earthy, s = sweet.

training and test set, respectively. Except for the green-earthy group the level of correct classification of the different aroma qualities for the training set is above 85.0%, for five of the sensory properties even all compounds are perfectly classified (see Table 3). The prediction of the aroma impression of the test compounds is presented in Table 4, showing that only four out of 30 test molecules are not correctly forecasted.

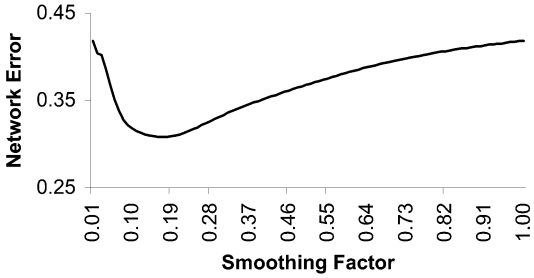

Figure 6. Error of the trained network in relationship to the smoothing constant σ.

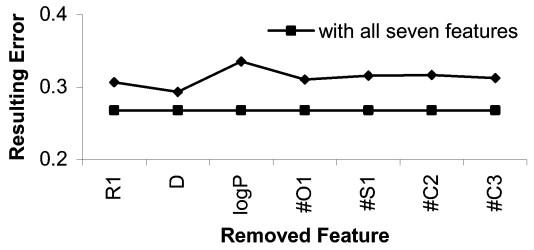

Figure 7. Sensitivity analysis for the input features of the GRNN. The notations of the descriptors are defined in Table 1.

General Regression Neural Network (GRNN). For GRNN the output for the five aroma impressions green, bellpepper, nutty, earthy, and sweet is coded in a binary way: an aroma impression is considered to be possessed if the value of the output neuron is higher than a threshold value of 0.5 ("presence"), while values below 0.5 point out that this aroma quality is missing ("absence"). As for PNN, 30 structures are selected from the data set of 133 pyrazines to build up a test set. Smoothing constants σ between 0.01 and 1.0 are tested in order to find the optimal one, which minimizes the error of the network, whereby the errors are defined as the sum of the squared differences between the predicted and actual output values on each output unit. As can be concluded from Figure 6, the smallest error is obtained with values for σ between 0.16 and 0.18. Furthermore, any σ in the range from 0.11 to 0.23 yields results only slightly worse than those for the best σ-values. The smoothing constant σ for the presented network is adjusted to a value of 0.18.

The best results are obtained with a network configuration consisting of 7 units in the input layer, 103 pattern neurons, 6 neurons in the summation layer, and 5 output units. The input neurons represent the values of the following molecular descriptors: logP, number of carbon atoms of substituents R_2 and R_3, number of sulfur and oxygen atoms of substituent R_1, charge of the first atom of substituent R_1, and the dipole moment of the structures. Figure 7 demonstrates the changes in the average absolute error by removing the input variables in turn. For the combination of all seven molecular characteristics the average absolute error for the prediction of the training set is 0.268. The most important descriptor contributing to the final network is logP, whose removal from the input variables results in an increased error with a value of 0.335.

The standard Pearson-R correlation coefficients between the target and predicted output values and the classification accuracy for the different aroma impressions are depicted in Table 5. Table 6 summarizes the performance of the GRNN model against the training set. From the total of 515 decisions (103 compounds combined with five aroma quali-

Table 5. Correlation Coefficients and Prediction Accuracy for the Different Aroma Impressions

	correlation coefficient		prediction accuracy [%]	
aroma quality	training	test	training	test
green	0.800	0.779	86.4	90.0
bellpepper	0.874	0.941	93.2	100.0
nutty	0.835	0.561	93.2	83.3
earthy	0.710	0.734	83.5	83.3
sweet	0.814	0.632	93.2	93.3

Table 6. Classification Performance of the Training Set for the Various Aroma Qualities by GRNN[a]

aroma quality	correct	wrong
green	89	14
bellpepper	96	7
nutty	96	7
earthy	86	17
sweet	96	7

[a] Correct = number of compounds which are correctly predicted to possess or not possess the defined aroma impression. Wrong = number of wrong decisions for the "presence" or "absence" of the given aroma quality.

ties) for the training set, 89.9% are correct. As can be expected from the lower correlation, the aroma quality earthy is more often misclassified than the other ones. On the other hand, the high correlation coefficient for bellpepper aroma results from the small number of wrong decisions. The prediction of the test set is presented in Table 4, whereby 90.0% of the aroma qualities of the test compounds are correctly predicted.

Multiple Linear Regression. Two linear regression models are calculated, whereby the encoding of the aroma impressions for these models is given in Table 2. Model MLR I corresponds to PNN. It differs between the same eight aroma impressions as the PNN model by using the same six descriptors as input variables and excluding the same 30 structures as external test set. The same is true for MLR II which is the linear opposite to GRNN (distinction between 18 aroma qualities by use of seven descriptors). Due to the use of the same test arrangement, comparison between the results of the linear and nonlinear methods is possible.

The resulting MLR models are presented in eqs 1 and 2, where AI denotes for the aroma impression of the studied compounds.

$$AI_{(MLR\ I)} = 3.98 + 1.81*R1 - 2.11*logP + 0.02*MS4 + 1.07*\#O1 + 0.31*\#C2 + 0.18*K3 \quad (1)$$

$$AI_{(MLR\ II)} = 9.02 + 1.52*R1 - 1.46*D - 2.21*logP + 2.20*\#O1 + 2.53*\#S1 + 0.05*\#C2 - 0.64*\#C3 \quad (2)$$

The quality of the two MLR models is judged by the correlation coefficient (r), the standard deviation (s), and the Fisher F-value (F). The statistics are depicted in Table 7.

As can be seen from the statistical parameters (low correlation coefficient (r), high standard deviation (s)), MLR does not work properly on the studied classification problem. Consequently, the prediction of the test set is not satisfying, as shown in Figures 8 and 9. Residues which are bigger than 0.5 characterize a misclassification into another than the literature-given aroma impression. As can be seen, most of the test compounds are, therefore, not correctly forecasted.

Table 7. Statistics of the MLR Studies

	MLR I	MLR II
r	0.57	0.55
s	1.88	3.69
F	6.72	5.96
nos[a]	120	133
noc[b]	6	7

[a] Number of studied structures. [b] Number of components.

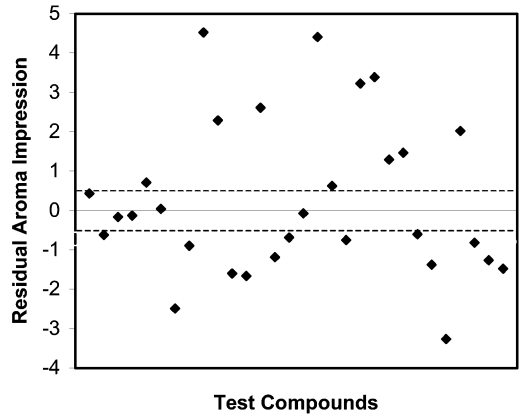

Figure 8. Residuals of the aroma quality prediction by MLR I.

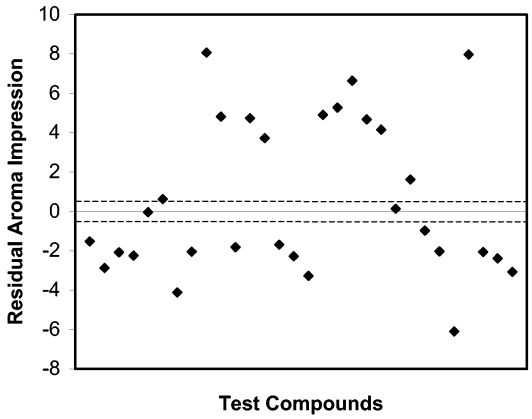

Figure 9. Residuals of the aroma quality prediction by MLR II.

DISCUSSION

Comparison of the obtained results from MLR and BNNs point out that the success of BNNs is due to the method and not to an efficient system of parameters. The obtained results from PNN show that this network type is able to distinguish between the aroma impressions green, bellpepper, and earthy as well as between the aroma mixtures green-nutty, green-earthy, bellpepper-nutty, bellpepper-earthy, and nutty-sweet with a correct prediction of 90.8% of the 120 pyrazines constituting the data set. This differentiation is enabled by combination of six molecular descriptors. From the sensitivity analysis it can be concluded that the networks' decision to which aroma class a single compound belongs is strongest influenced by the number of carbon atoms of the substituent R_2. Most of the studied aroma impressions show side-chains at this position of a length between two and five carbon atoms. The nutty-sweet smelling pyrazines contain only up to two carbon atoms at this substituent and can, therefore, be separated from the other sensory properties by this descriptor. Furthermore, a slight differentiation of the bell-pepper and the green-earthy group from all the other aroma impressions is possible, as they show bulky groups at

substituent R_2. Additional information for the differentiation is obtained from the logP-values of the structures, which allows for identification of the pyrazines belonging to the earthy, green-nutty, or nutty-sweet odor as they preferentially have a significantly lower logP-value than the other impressions. Moreover, the number of oxygen atoms at position R_1 is of importance for the classification problem. While the qualities earthy, green-nutty, and green-earthy contain no oxygen atom, bellpepper-nutty and bellpepper-earthy smelling pyrazines do. Differences can also be seen from the charge of the first atom of the substituent R_1. The green-earthy group shows positive values for this descriptor, the earthy quality less negative charges than all the other impressions. The classification of the eight aroma impressions is additionally influenced by the molecular surface of substituent R_4. As green, bellpepper, green-nutty, green-earthy, and nutty-sweet smelling pyrazines normally contain only a hydrogen at position R_4, their molecular surface at this position is rather small. In comparison to this the earthy, bellpepper-nutty, and bellpepper-earthy groups possess more bulky substituents, which results in higher values for their molecular surface at substituent R_4. Very little information only can be obtained from the Kappa shape index of third order, which encodes information about the centrality of branching. This descriptor allows to identify the earthy smelling group, as these structures generally show lower values for the Kappa shape index as the other aroma impressions.

In comparison to the PNN, the GRNN is able to predict the presence or absence of distinct aroma impressions with the good accuracy of 89.9% of all decisions. Each of the aroma qualities green, bellpepper, nutty, earthy, and sweet is characterized by several features. While the descriptors logP, charge of the first atom of substituent R_1, and dipole moment of the whole structure do not allow to draw clear conclusions, the counts for various atom types give small insight into the structural differences of the five aroma impressions. Structures containing a sulfur atom at position R_1 and a long side-chain at substituent R_3 preferentially possess a green odor. On the other hand, the occurrence of an oxygen atom at position R_1 and a bulky substituent at position R_2 denotes for the bellpepper group. Furthermore, pyrazines with bellpepper odor contain a short side-chain at substituent R_3, a characteristic which they share with the nutty and earthy smelling structures. The nutty group is additionally indicated by a short side-chain at substituent R_2. This nonbulky substituent is also common for the sweet aroma impression, which is, moreover, marked by an oxygen atom at position R_1. As can be seen from these conclusions, some of the mentioned characteristic features are shown by multiple aroma qualities, and, therefore, a clear distinction between the various groups is only possible by a combination of all seven descriptors. Furthermore, some structures contain the typical attributes of more than one aroma quality, which results in a mixed aroma impression.

CONCLUSION

Important structural criterions, which turned out to be characteristic for the five "pure" aroma qualities bellpepper, green, nutty, earthy, and sweet are summarized in Figure 10. The continuous line represents structural features obtained by interpretation of the PNN results, while the dashed line marks characteristics outlined by GRNN. The meanings of the used symbols can be derived from Table 8. It can be seen that the information obtained from the two different network types is partly the same. For example both networks point out the requirement of a long side-chain at position R_2 for the bellpepper aroma impression. However, in principle the trained networks use diverse input variables and, as a consequence, allow for gaining information on different structural features of the aroma classes. Therefore, the combination of the results of the PNN with those of the GRNN gives the most complete information about structural differences between the various aroma qualities.

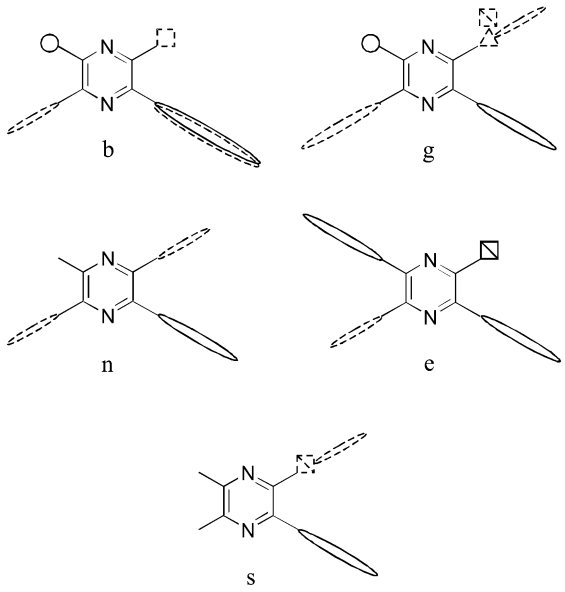

Figure 10. Structural characteristics of the studied aroma qualities (g = green, b = bellpepper, n = nutty, e = earthy, s = sweet). The continuous line represents results obtained from PNN, while the broken line stands for results from GRNN. The meanings of the different symbols are defined in Table 8.

Table 8. Meaning of the Symbols Representing the Important Structural Criterions Shown in Figures 10 and 11

Symbol	Notation
	side-chain with more than six carbon atoms
	side-chain with up to six carbon atoms
	side-chain with up to two carbon atoms
□	mostly an oxygen atom at this position
⧅	no oxygen atom at all at this position
△	mostly a sulfur atom at this position
○	usually a hydrogen atom at this position

As an example for the structural requirements of the mixed aroma impressions, Figure 11 represents the typical features of the green-nutty and bellpepper-earthy smelling groups. Closer inspection and comparison of the structural criterions points out that the mixed aroma impressions really show characteristics of both "pure" aroma impressions they are composed of. The requirements for the green-nutty group constitute a side-chain at position R_2 which should not be too long. This is a common structural feature among the green and nutty aroma class. Moreover, the mixed aroma impression shows two further characteristics of the green

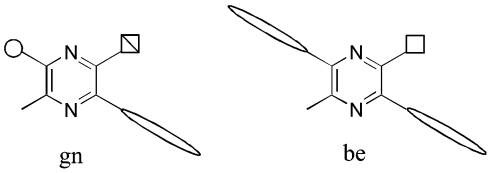

Figure 11. Typical features of the mixed aroma impressions green-nutty (gn) and bellpepper-earthy (be). The meanings of the different symbols are defined in Table 8.

smelling pyrazines, namely a hydrogen atom at position R_4 and the missing of an oxygen atom at position R_1. The structures of the bellpepper-earthy smelling group share the oxygen atom at position R_1 with the bellpepper aroma quality, while the length of the two side-chains R_2 and R_4 is the same as in the case of earthy smelling pyrazines. The combination of these facts results in the mixed aroma impression of bellpepper and earthy quality.

From these examples it can be concluded that the effectiveness and the reliability of BNNs to multiple-category problems is rather high. Both, namely PNN as well as GRNN, have of course their advantages and disadvantages. The results of the PNN are relatively easy to interpret. However, the studied aroma classes need to be well-defined, which means that some of the aroma impressions cannot be investigated due to the lack of structures showing these qualities. For GRNN, all possible mixtures of aroma impressions can be studied, but, on the other hand, the evaluation and interpretation of the obtained results are much more complicated than in the case of a PNN. However, the results of our study provide further proof that both types of BNNs are well suited to deal with many different types of aroma impressions, including aroma mixtures, and that these methods could be extended to the study of other sets of aroma compounds to obtain information about their structural differences.

Supporting Information Available: Structures and investigated aroma impressions of all studied compounds (Table 1) and experimental aroma impressions of all studied molecules (Table 2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Flower, D. R. The lipocalin protein family: structure and function. *Biochemical 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.; Vincent, F.; Spinelli, S.; Campanacci, V.; Grolli, S.; Cambillau, C. Mammalian odorant binding proteins. *Biochim. Biophys. Acta* **2000**, *1482*, 229−240.
(7) Buck, L.; Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **1991**, *65*, 175−187.
(8) Malnic, B.; Hirono, J.; Sato, T.; Buck, L. B. Combinatorial Receptor Codes for Odors. *Cell* **1999**, *96*, 713−723.
(9) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(10) Wailzer, B.; Klocker, J.; Buchbauer, G.; Ecker, G.; Wolschann, P. Prediction of the Aroma Quality and the Threshold Values of some Pyrazines using Artificial Neural Networks. *J. Med. Chem.* **2001**, *17*, 2805−2813.
(11) Seifert, R. M.; Buttery, R. G.; Guadagni, D. G.; Black, D. R.; Harris; J. G. Synthesis of some 2-Methoxy-3-Alkylpyrazines with Strong Bellpepper-Like Odors. *J. Agric. Food Chem.* **1970**, *18*, 246−249.
(12) Parliment, T. H.; Epstein, M. F. Organoleptic Properties of some alkyl-substituted alkoxy- and alkylthiopyrazines. *J. Agric. Food Chem*. **1973**, *21*, 714−716.
(13) Pittet, A. O.; Hruza, D. E. Comparative Study of Flavor Properties of Thiazole Derivatives. *J. Agric. Food Chem*. **1974**, *22*, 264−269.
(14) 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.
(15) Masuda, H.; Mihara, S. Synthesis of Alkoxy-, (Alkylthio)-, Phenoxy-, and (Phenylthio)pyrazines and Their Olfactive Properties. *J. Agric. Food Chem*. **1986**, *34*, 377−381.
(16) Shibamoto, T. Odor Threshold of Some Pyrazines. *J. Food Sci.* **1986**, *51*, 1098−1099.
(17) Masuda, H.; Mihara, S. Olfactive Properties of Alkylpyrazines and 3-Substituted 2-Alkylpyrazines. *J. Agric. Food Chem*. **1988**, *36*, 584−587.
(18) Mihara, S.; Masuda, H. Structure-Odor Relationships for Disubstituted Pyrazines. *J. Agric. Food Chem*. **1988**, *36*, 1242−1247.
(19) 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.
(20) Boelens, M. H.; Van Gemert, L. J. Structure−Activity Relationships of Natural Volatile Nitrogen Compounds. *Perfum. Flavour* **1995**, *20*, 63−76.
(21) Grosch, W.; Wagner, R.; Czerny, M.; Bielohradsky, J. Structure-odor activity relationships of earthy smelling alkylpyrazines. *Z. Lebensm. Unters. Forsch.* **1999**, *208*, 308−316.
(22) Hyperchem 5.0; Hypercube Inc.: Gainesville, FL, 1997.
(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. GAUSSIAN 98, Revision A.6; Gaussian, Inc.: Pittsburgh, PA, 1998.
(24) TSAR 3.2; Oxford Molecular, Ltd.: Oxford, England, 1999.
(25) Trajan Neural Networks 4.0; Trajan Software Ltd.: Durham, UK, 1999.
(26) Specht, D. F. Probabilistic Neural Networks. *Neural Networks* **1990**, *3*, 109−118.
(27) Specht, D. F. A General Regression Neural Network. *IEEE Trans. Neural Networks* **1991**, *2(6)*, 568−576.
(28) Yang, Z. R.; Platt, M. B.; Platt, H. D. Probabilistic Neural Networks in Bankruptcy Prediction. *J. Business Res.* **1999**, *44*, 67−74.
(29) Simon, L.; Nazmul Karim, M. Probabilistic neural networks using Bayesian decision strategies and a modified Gompertz model for growth phase classification in the batch culture of *Bacillus subtilis*. *Biochem. Eng. J.* **2001**, *7*, 41−48.
(30) Hansen, J. V.; Meservy, R. D. Learning experiments with genetic optimization of a generalized regression neural network. *Decision Support Systems* **1996**, *18*, 317−325.
(31) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.

CI0202640