Sensors and Actuators B 255 (2018) 508–518

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

journal homepage: www.elsevier.com/locate/snb

Odorant clustering based on molecular parameter-feature extraction and imaging analysis of olfactory bulb odor maps

Liang Shang[a], Chuanjun Liu[a,b], Yoichi Tomiura[c], Kenshi Hayashi[a,*]

[a] Department of Electronics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan
[b] Research Laboratory, U.S.E. Co., Ltd., Tokyo 150-0013, Japan
[c] Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan

ARTICLE INFO

Article history:
Received 8 February 2017
Received in revised form 11 July 2017
Accepted 2 August 2017
Available online 16 August 2017

Keywords:
Olfactory bulb odor map
Image analysis
Physicochemical parameter
t-Distributed stochastic neighbor embedding
Classification analysis
Artificial neural network

ABSTRACT

Progress in the molecular biology of olfaction has revealed a close relationship between the structural features of odorants and the response patterns they elicit in the olfactory bulb. Molecular feature-related response patterns, termed odor maps (OMs), may represent information related to basic odor quality. Thus, studying the relationship between OMs and the molecular features of odorants is helpful for better understanding the relationships between odorant structure and odor. Here, we explored the correlation between OMs and the molecular parameters (MPs) of odorants by taking OMs from rat olfactory bulbs and extracting feature profiles of the corresponding odorant molecules. 178 images of glomerular activities in olfactory bulb that are corresponding to odorants were taken from the OdorMapDB, a publicly accessible database. The gray value of each pixel was extracted from the images (178 × 357 pixels) to fabricate an image matrix for each odorant. Forty-six molecular feature parameters were calculated using BioChem3D software, which was used to construct a second matrix for each odorant. Correlation analysis between the two matrixes was first carried out by establishing coefficient maps. Results from hierarchical clustering showed that all parameters could be segregated into seven clusters, and each cluster showed a relatively similar response pattern in the olfactory bulb. Using the information from the OMs and MPs, we mapped odorants in 2D space by incorporating dimension-reducing techniques based on principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). Artificial neural network models based on the OM and MP feature values were proposed as a means to identify odorant functional groups. An OM-PCA-based model calibrated via extreme learning machine (ELM) was 94.81% and 93.02% accuracy for the calibration and validation sets, respectively. Similarly, an MP-t-SNE-based model calibrated by ELM was 86.67% and 93.35% accuracy for the calibration set and the validation set, respectively. Thus, this research supports a structure-odor relationship from a data-analysis perspective.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

As the most primitive sense, olfaction plays an essential role in our daily lives [1]. Over 400,000 chemicals are known to produce a sense of smell in people. Recent research has reported that we can discriminate more than 1 trillion olfactory stimuli [2]. The mechanism through which olfactory perception is achieved remained basically unknown until Richard and Buck discovered odorant receptors and described the organization of the olfactory system [3,4]. Since then, our knowledge regarding olfactory perception, particularly at the molecular level, has grown significantly [5–7].

Research into the response patterns of neurons in the olfactory bulb (OB) has helped clarify the mechanism of biological olfaction [8]. The first relay station of the olfactory system is the olfactory bulb (OB), which has a cortical structure with distinct layers and numerous glomerular modules [9]. Molecular features of odorants have been shown to be represented by spatiotemporal patterns of activity across olfactory sensory neurons in the OB [10,11]. Odorants are discriminated and recognized in mammal brains through analysis of these glomerular activity patterns [12,13]. By imaging the 2-deoxyglucose uptake in rat glomeruli, Johnson and his team systematically mapped spatially odorant-evoked activity into two dimensional (2D) images for more than 300 odorants [14], and these odor maps now comprise a database (OdorMapDB) that can be freely accessed (http://gara.bio.uci.edu/) [15]. Johnson et al.

* Corresponding author.
E-mail addresses: shang.liang.100@s.kyushu-u.ac.jp (L. Shang), liu@o.ed.kyushu-u.ac.jp (C. Liu), tom@inf.kyushu-u.ac.jp (Y. Tomiura), hayashi@ed.kyushu-u.ac.jp (K. Hayashi).

http://dx.doi.org/10.1016/j.snb.2017.08.024
0925-4005/© 2017 Elsevier B.V. All rights reserved.

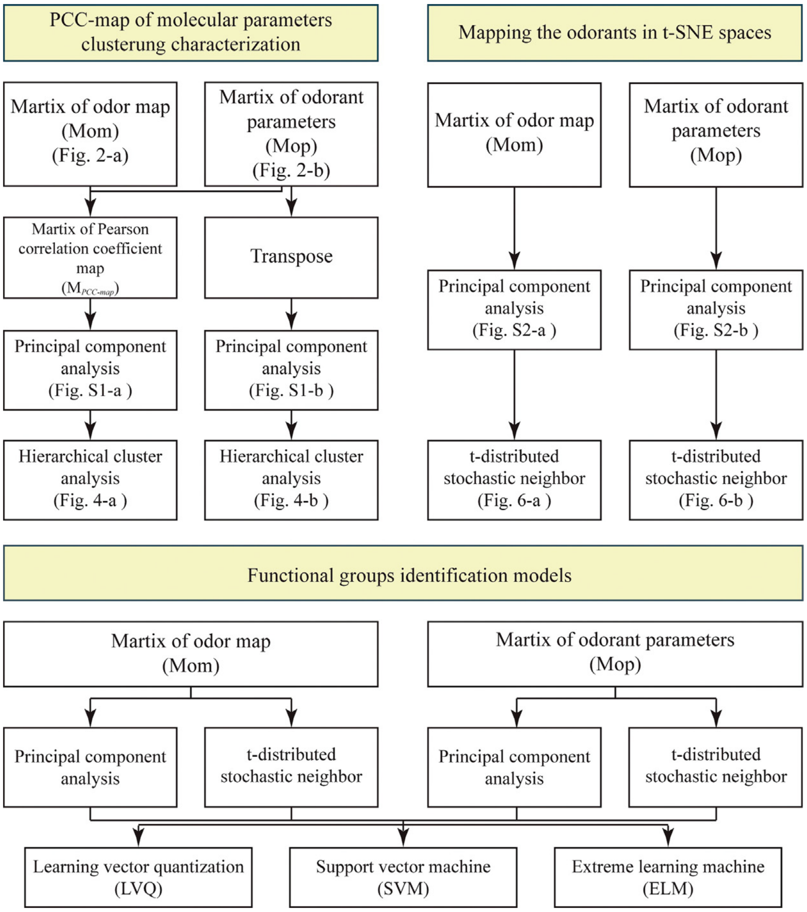
PCC-map of molecular parameters clusterung characterization
Martix of odor map (Mom) (Fig. 2-a)
Martix of odorant parameters (Mop) (Fig. 2-b)
Martix of Pearson correlation coefficient map ($M_{PCC-map}$)
Transpose
Principal component analysis (Fig. S1-a)
Principal component analysis (Fig. S1-b)
Hierarchical cluster analysis (Fig. 4-a)
Hierarchical cluster analysis (Fig. 4-b)
Mapping the odorants in t-SNE spaces
Martix of odor map (Mom)
Martix of odorant parameters (Mop)
Principal component analysis (Fig. S2-a)
Principal component analysis (Fig. S2-b)
t-distributed stochastic neighbor (Fig. 6-a)
t-distributed stochastic neighbor (Fig. 6-b)
Functional groups identification models
Martix of odor map (Mom)
Martix of odorant parameters (Mop)
Principal component analysis
t-distributed stochastic neighbor
Principal component analysis
t-distributed stochastic neighbor
Learning vector quantization (LVQ)
Support vector machine (SVM)
Extreme learning machine (ELM)

Fig. 1. Schematic diagram of data processing.

concluded that clustering responses on the glomerular surface to the molecular features of odorants is likely a general strategy for odor encoding [16]. Additionally, Mori et al. have summarized nine molecular-feature clusters that they found at stereotypical OB positions [17].

Research into the molecular biology of olfactory perception has revealed a close relationship between the structural features of odorants and their olfactory perception. For example, functional groups and carbon-chain length play central roles in odor perception [18–20]. Indeed, several studies have been devoted to describing the structure-odor relationship [21,22]. For measuring the similarity between two odorants, a vector containing 1664 descriptors is applied to describe the structure or shape of molecules, and the physicochemical space (principal component space) is used to evaluate the “distance” between them [23,24]. Further, mass spectra and infrared absorption spectra are used to encode odors via artificial neural networks or self-organizing maps [25]. However, olfaction is extremely complex, and a complete understanding of the structure-odor relationship has yet to be realized. Despite efforts have been made to measure smell, none can describe all pertinent aspects of olfactory perception [26], and the results are difficult to explain without knowledge of biology [27].

Finding basic odors is difficult because the numbers of olfactory receptors and odorants are very large. While most studies focus on attempting to connect odorant physicochemical properties to olfactory perception [28], objective information such as OB response patterns has rarely been considered [29]. Nevertheless, studying the relationship between olfactory response patterns and the structural features of odorants can be helpful in understanding the mechanisms underlying olfactory perception and for predicting the structure-odor relationship [30]. Consequently, the primary goal of this experiment was to explore the relationships between odor-induced patterns of activity and the associated molecular features.

We first obtained 2-deoxyglucose glomerular activity-pattern images for 178 odorants from the Johnson freely available odor-map database [31]. For each map, the gray value of each pixel was extracted from the images to fabricate a 178 × 70329 image matrix. Forty-six molecular feature parameters for the odorants were calculated using BioChem3D software. A schematic of the data-processing method is shown in Fig. 1. Based on the characteristic variables extracted by principal component analysis (PCA), hierarchical clustering analysis (HCA) was performed on the Pearson correlation coefficient maps (PCC-maps) to investigate the effects of the molecular parameters. 2D artificial cluster maps based on the olfactory and molecular information were generated via t-distributed stochastic neighbor-embedding (t-SNE). Based on these datasets, three machine learning models—learning vector quantization (LVQ) network, support vector machine (SVM), and extreme learning machine (ELM)—were employed to establish odorant function-group discrimination models. We then assessed the feasibility of odor maps (OMs) and molecular parameters (MPs) for odorant function-group classification using each model.

2. Materials and methods

2.1. Glomerular activity patterns and molecular parameters

We used glomerular activity patterns (odor maps) from the dorsal part of rat OB. OMs (grey image), chemical abstracts service

Table 1
46 types of molecular parameters extracted by ChemBio 3D.

No.	Molecular parameter	No.	Molecular parameter	No.	Molecular parameter
1	Boiling point	18	Total energy	35	Shape coefficient
2	Critical pressure	19	Dipole	36	Sum of degrees
3	Critical temperature	20	Number of Hbond acceptors	37	Sum of valence degrees
4	Critical volume	21	Number of Hbond Donors	38	Topological diameter
5	Gibbs free energy	22	Ovality	39	Total connectivity
6	Heat of formation	23	Principal moment	40	Total valence connectivity
7	Henry's law constant	24	Elemental analysis	41	Wiener index
8	Ideal gas thermal capacity	25	Molecular weight	42	Core-core repulsion
9	LogP	26	LogS	43	COSMO area
10	Melting point	27	Pka	44	COSMO volume
11	Mol refractivity	28	Balaban index	45	Electronic energy
12	Vapor pressure	29	Cluster count	46	Ionization potential
13	Water solubility	30	Molecular topological index		
14	Connolly accessible area	31	Num rotatable bonds		
15	Connolly molecular area	32	Polar surface area		
16	Connolly solvent excluded volume	33	Radius		
17	Exact mass	34	Shape attribute		

(CAS) numbers, and functional group labels were extracted using semi-automatic and manual methods from the Johnson and Leon database. Forty-six MPs for 178 odorants (Table S1) were determined using the MOPAC and GAMESS packages in ChemBio3D Ultra 11.0 (2008, Cambridge Soft, Massachusetts, USA) by establishing 3D models for odorants based on simplified strings of molecular input-line entry specifications (SMILES). All the parameters used in this study are listed in Table 1.

2.2. Construction of PCC-maps for molecular features

We used the following computational process to generate PCC-maps for investigating the relationship between OMs and MPs.

Step 1. Each 357 × 197 OM was transformed into a 1 × 70329 vector. We created a matrix of odor maps (Mom) by combining the 178 OMs (gray images) (Fig. 2a).

$$Mom = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,70329} \\ P_{2,1} & P_{2,1} & \cdots & P_{2,70329} \\ \vdots & \vdots & \vdots & \vdots \\ P_{178,1} & P_{178,2} & \cdots & P_{178,70329} \end{bmatrix}_{178\times70329} \quad (1)$$

Step 2. We created a similar matrix of odorant parameters (Mop), which contained 46 molecular parameters from 178 odorants (Fig. 2b).

$$Mop = \begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,46} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,46} \\ \vdots & \vdots & \vdots & \vdots \\ V_{178,1} & V_{178,2} & \cdots & V_{178,46} \end{bmatrix}_{178\times46} \quad (2)$$

Step 3. We calculated the correlation coefficients between molecular parameters and the gray-level value of each OM pixel. The Pearson correlation-coefficient (PCC) matrix $M_{PCC-map}$ was defined as follows:

$M_{PCC-map} = \{R_{i,j} | R_{i,j} = Cor(Mom(i, :), Mop(j, :)), i = 1, 2, 3, \ldots, 49, j = 1, 2, 3, \ldots, 70329\}$where, $Mom(i, :)$ and $Mom(j, :)$ indicate the i-th and j-th row vector in Mom and Mop, respectively. $Cor(x, y)$ was defined in Formula (3).

$$Cor(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (3)$$

where, $\bar{x}$ and $\bar{y}$ are the mean values of vector $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. N is the dimension of vector $\boldsymbol{x}$ or $\boldsymbol{y}$ (here, N was 70329). Thus, the PCC matrix $M_{PCC-map}$ was established as Formula (4).

$$M_{PCC-map} = \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,70329} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,70329} \\ \vdots & \vdots & \vdots & \vdots \\ R_{49,1} & R_{49,2} & \cdots & R_{49,70329} \end{bmatrix}_{49\times70329} \quad (4)$$

Step 4. We performed HCA based on the Euclidean distances in the latent variables extracted by PCA, and applied Ward's method as a similarity criterion to cluster the 46 molecular parameters into homogeneous groups. The clustering results were then evaluated and analyzed to investigate the relationships between MPs and OMs.

Step 5. We reshaped each row vector of $M_{PCC-map}$ as a 357 × 197 matrix and obtained the PCC-map for each MP.

2.3. Selection methods for characteristic variables

2.3.1. Principal component analysis (PCA)

PCA is generally used to extract characteristic variables from a high-dimensional data set [32]. PCA can remove linear and duplicated information by constituting principal components (PCs) from original data. The PCs listed first in the output are selected as characteristic variables according to the cumulative contribution rate, while those listed at the end of the output are removed because of noise [33].

2.3.2. Barnes-Hut t-distributed stochastic neighbor embedding (t-SNE)

t-SNE is a novel, unsupervised embedding method that has been used to visualize high-dimensional data at a lower dimension [34], and that allows dataset embeddings to be learned. The computational process is as follows.

Step 1. Given a high dimension dataset $L = \{x_i | x_i \in R^m, i = 1, 2, \ldots, N\}$, where x_i is a vector $(1 \times m)$ of the i-th sample, m indicates the total variable number of each sample, and N indicates the total

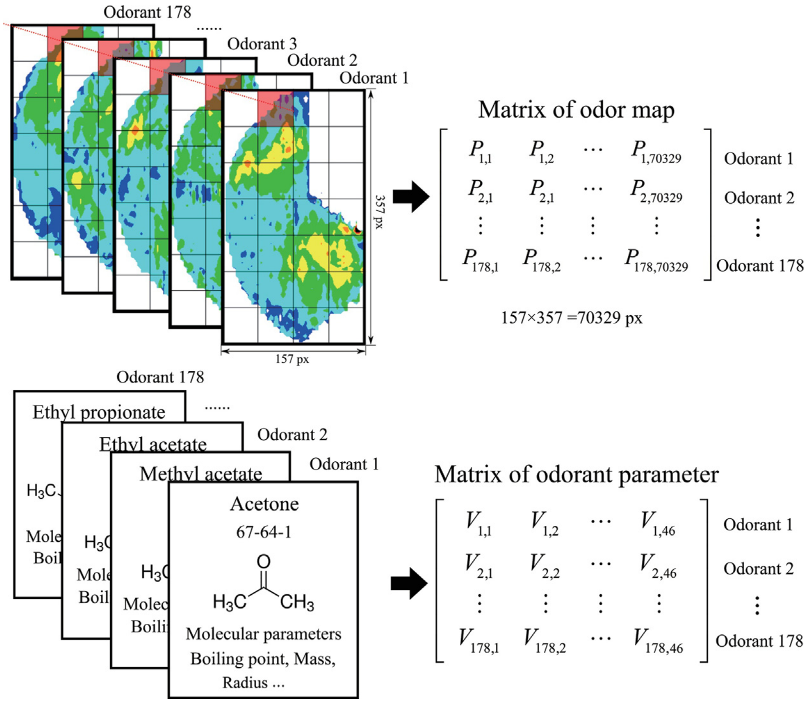
Odorant 178
......
Odorant 3
Odorant 2
Odorant 1
357 px
157 px
Matrix of odor map
$P_{1,1}$ $P_{1,2}$ ⋯ $P_{1,70329}$ Odorant 1
$P_{2,1}$ $P_{2,1}$ ⋯ $P_{2,70329}$ Odorant 2
$P_{178,1}$ $P_{178,2}$ ⋯ $P_{178,70329}$ Odorant 178
157×357 =70329 px
Odorant 178
Ethyl propionate
......
Odorant 2
Ethyl acetate
Odorant 1
Methyl acetate
Acetone
67-64-1
O
H_3C CH_3
Molecular parameters
Boiling point, Mass,
Radius ...
Matrix of odorant parameter
$V_{1,1}$ $V_{1,2}$ ⋯ $V_{1,46}$ Odorant 1
$V_{2,1}$ $V_{2,2}$ ⋯ $V_{2,46}$ Odorant 2
$V_{178,1}$ $V_{178,2}$ ⋯ $V_{178,46}$ Odorant 178

Fig. 2. Schematic diagram of obtaining matrix of odor maps (Mom) (a) and molecular parameters (Mop) (b).

sample number in the dataset, the function $d(x_i, x_j)$ is the computed distance between a pair of samples. Here, we calculated the Euclidean distance, $d(x_i, x_j) = \|x_i - x\|$.

Step 2. We used Formula (5) to calculate the pairwise similarities (p_{ij}) between the samples.

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\delta^2)}{\sum\ k \neq l(-\|x_l - x_k\|^2/2\delta^2)} \qquad (5)$$

where δ is the variance parameter of the Gaussian function.

Step 3. We calculated the similarity (q_{ij}) between target values, $T = \{y_i | y_i \in R^m, i = 1, 2, \ldots, N\}$, in a low-dimensional space ($m = 2$ or 3) using a normalized Student-t kernel with one degree of freedom.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \qquad (6)$$

Step 4. Based on Kullback-Leibler divergence measuring, we determined the locations of the embedding points (Y).

$$C(Y) = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (7)$$

The optimal low-dimensional representation Y was calculated by minimizing $C(Y)$. Barnes-Hut t-SNE was employed to establish 2D artificial odor maps. Here, t-SNE was calculated using the R (version 3.2.2) package named "Rtsne" (version 0.1). More detailed explanations of t-SNE processing can be found elsewhere [35].

2.4. Sample division method

Rational division of sample sets is crucial for improving the validation accuracy of models [36]. The calibration set should include the utmost main information from the original samples. Another benefit of rational sample-set division is that it avoids overlapping in machine-learning models. Here, the Kennard-Stone (KS) algorithm was selected for the sample partition [37]. The details of this process have been described by other research [38,39]. In the current study, 178 odorants were divided into calibration and validation sets via the KS method. The ratio of samples between calibration and validation sets was 3–1. The calibration set therefore contained 135 samples and the validation set contained 43.

2.5. Modeling methods

2.5.1. Learning vector quantization (LVQ)

As a supervised learning algorithm for classification, LVQ is mostly applied for pattern recognition or qualitative analysis based on self-organizing maps [40]. An LVQ network can be optimized by confirming the decision boundaries between neighboring groups. The network contains three layers: an input layer, a competitive layer, and a linear output layer [41]. In this study, LVQ1 was applied to establish the classification models. Additional information about this type of LVQ network can found elsewhere [42].

2.5.2. Support vector machine (SVM)

SVM is a powerful classification model based on statistical learning theory, which has been widely applied in machine vision, image processing, and pattern recognition [43,44]. By establishing a hyperplane as a decision surface, the positive-examples and counter-examples can be divided such that they are separated by the greatest possible distance [45]. Details regarding SVM have been published elsewhere [46,47]. In the current study, a radial basis function (RBF) was selected as the kernel function for the SVM model, which was established using the Libsvm (version 2.81) package [48].

2.5.3. Extreme learning machine (ELM)

ELM is an efficient single-hidden-layer feed-forward neural network that is widely used for establishing non-linear relationships because of its good performance at generalization [49]. ELM can also overcome some difficulties in traditional learning methods, such as learning rate and epochs [50]. If the number of hidden layer nodes is assigned, the weights between input neurons and hidden neurons can be chosen and fixed randomly [51]. Details regarding the computational process for ELM can be found elsewhere [52].

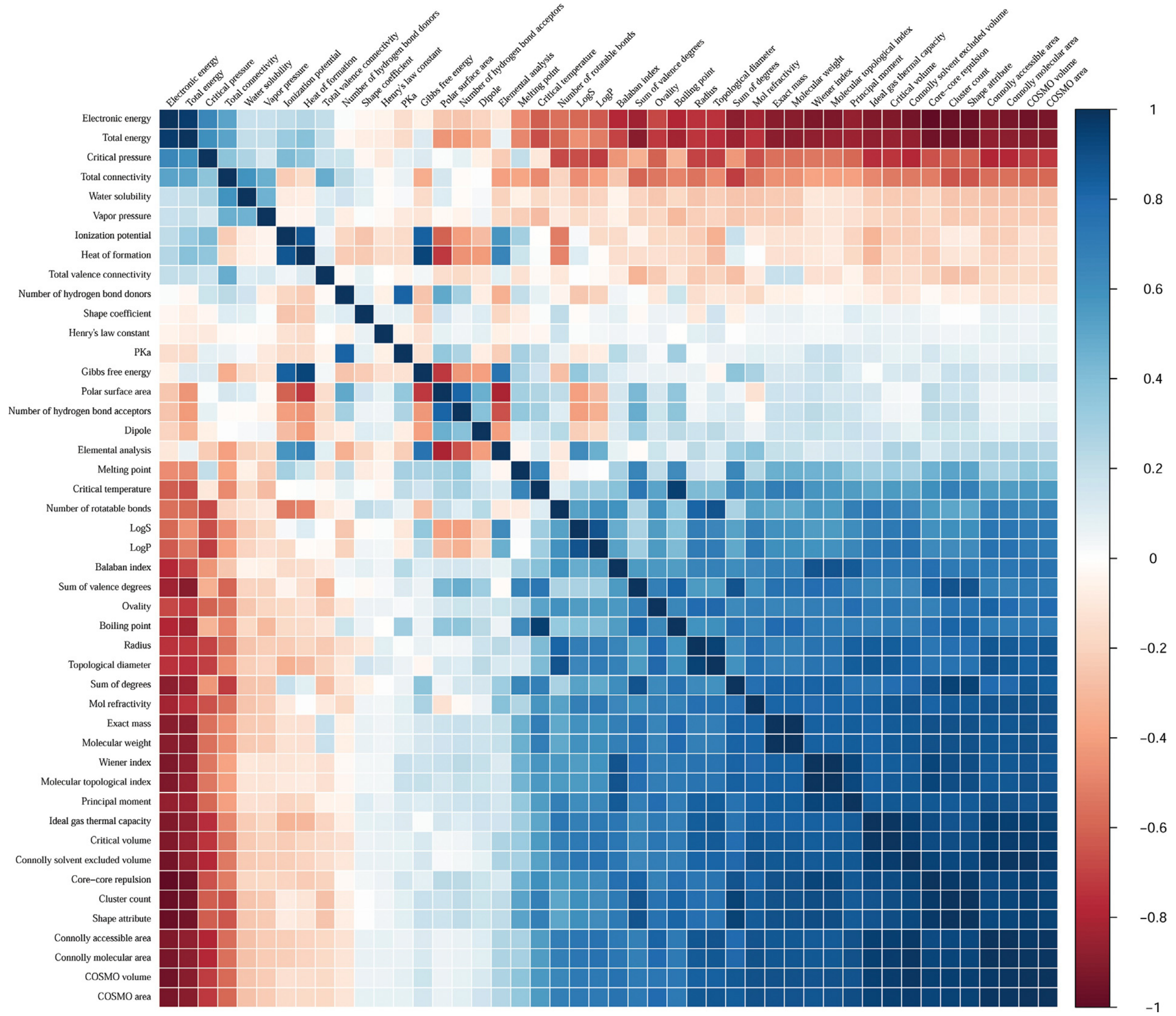

Fig. 3. Correlation map for 46 types of molecular parameters.

3. Results and discussion

3.1. Molecular parameters and functional group labels

In our previous study, PCA was carried out on the basis of 79 molecular parameters of odorants from the odor-map database [53]. The result indicated that the number of parameters that are important for generating odor maps was not large. Indeed, only 15 parameters showed strong relevance to the first 6 PCs. Based on the above results and limitation of the software (Chembio 3D), in this study we only calculate 46 parameters to carry out the analysis.

The research of Mori reveal that in the rat OB there exist nine independent zones which is corresponding to different functional groups [54]. Other evidence has shown that functional groups affect odor sensation. For example, dorants with the functional group '-COOH' are perceived as smelling like sweat [55]. The above results indicate that functional groups may be a good labels for odorant classification, and the identification of functional groups is a better way to understand odorant sensation. We summarized the functional group label that appeared in the OdorMapDB. The 300 odorants are labeled by seven functional groups in which most show crossed information, especially for complicated odorants with high molecular weights. For odorant classification, the labels should not intersect each other. To simplify the model and to improve prediction accuracy, 14 non-intersection labels collected from 178 odorants were chosen to test our hypothesis (Table S2).

3.2. Clustering characterization for PCC-maps of molecular parameters

A correlation heat map for the 46 molecular parameters (Fig. 3) shows that some parameters are linearly related and that some redundant information is included in the molecular parameter matrix. Therefore, before clustering, we performed PCA to reduce the dimensions of the PCC-maps and MPs. Based on Euclidean distances between the first four latent variables extracted by PCA (accumulative contribution = 90.9%, Fig. S1 a), HCA was performed to investigate the relation between MPs and OMs. The results wereas organized and depicted by a heat map shown in Fig. 4a. The horizontal dendrogram of the heat map shows the 46 MPs are clustered into 7 groups. All the PCC-maps were provided in Fig. 5. Similar response pattern is shown in each group. It indicated that the MPs in the same group could contribute the similar information to OMs. For example, most of parameters contained energy information are clustered in group 1, and parameters contained polarity information are clustered in group 2. Further, low correlation coefficients are observed for each molecular feature. This finding indicates that the relationships are non-linear, and that one odor receptor could be linked to multiple physicochemical odor-

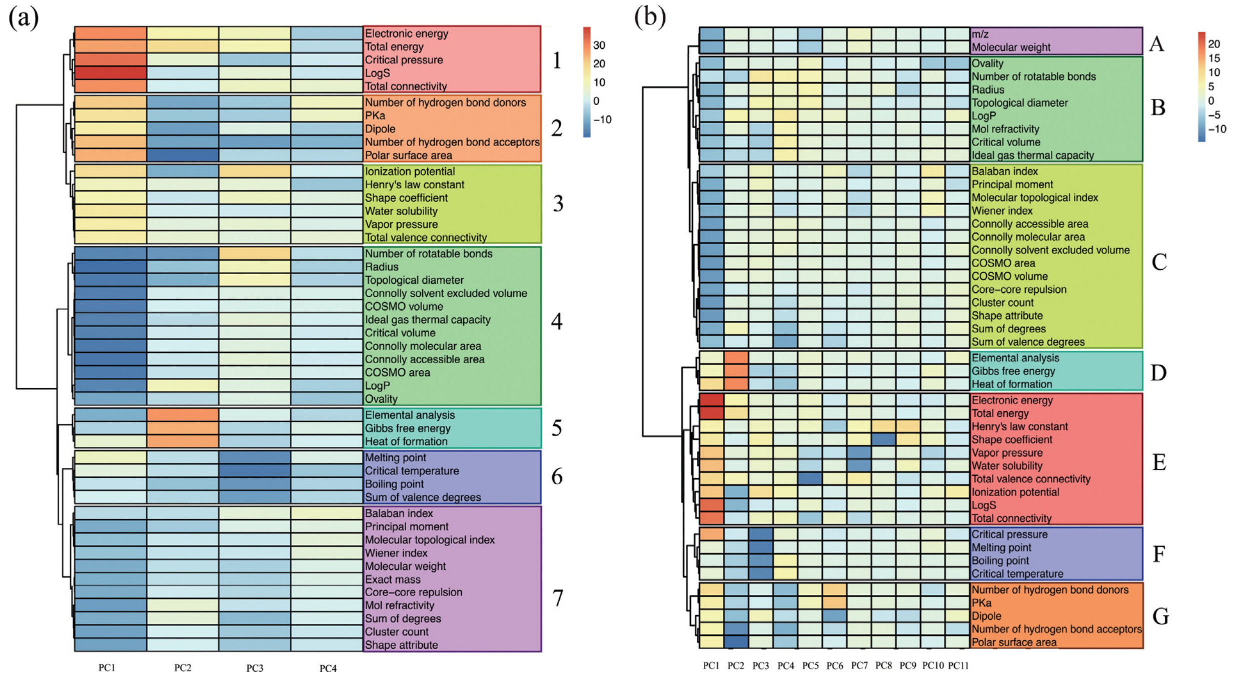

Fig. 4. Heat map and hierarchical dendrograms of the PCC-maps (a) and parameters (b) for 46 molecular parameters. Cluster analysis was performed by Ward's method on Euclidean distance of the first 4 PCs for R-maps. Each row indicated one type of molecular parameter, and each column indicated a PC.

ant features. These conclusions are supported by Kaeppler's and Johnson's work [28,56].

Through the same procedure, HCA was applied to the first 11 PCs (cumulative contribution rate = 91.4%) of the Mop (Fig. S1b), and the result is shown in Fig. 4b. Groups visualized by these heat maps shared some similarities to the PCC maps, such as cluster 2 and cluster G, cluster 5 and cluster D, and cluster 6 and cluster F. However, some parameters are clustered differently between the two heat maps. This indicates that these parameters are sensitive to olfactory information.

3.3. *Mapping the odorants in t-SNE space*

Next, we investigated the possibility of mapping odorants in 2D space. Before using t-SNE, we always employed PCA to extract vital information from the original matrixes. Generally, when PCs have more than 85% cumulative contribution from the original dataset, these PCs can be used to replace the originals [57]. Here, the first 80 PCs (cumulative contribution rate = 86.0%) for Mom and the first 23 PCs for Mop (cumulative contribution rate = 99.0%) were used as the inputs to the t-SNE analysis (Fig. S2). Next, the Barnes-Hut t-SNE algorithm was utilized to plot the odor maps in 2D space. In total, 178 odorants from odors with 14 functional groups were mapped into 2D t-SNE space. The initial dimensions were set to 80 for Mom and 23 for Mop, and the perplexity and maximum number of iterations were set to 50 and 10000, respectively. Table S1 shows the details for the data calculated by t-SNE.

The artificial map generated from the olfactory information is shown in Fig. 6a. It indicates that 64.04% of samples within the same category are clustered together (Table S3). Additionally, samples from the chemical categories *small aliphatic ester* ($C \leq 8$), *aliphatic or alicyclic hydrocarbon*, *aromatic*, and *carboxylic acid* are clustered in multiple groups. This indicates that carbon-chain length is a vital factor for *aliphatic esters* (cluster 1) and *alcohols* (cluster 9), and that branched chains play a role in distinguishing odorants with *hydrocarbon alkanes* (cluster 6) from those with *carboxylic acids* (cluster 12). Most odorants classified as *aromatic* (cluster 3, 4, and 7) are mapped in the left of t-SNE space. However, some *aromatic* odorants are scattered and not clustered together. We attribute this result to insufficient numbers of samples which cannot completely identify hidden patterns of the molecular structure. Some clusters, such as clusters 4, 7, 6-1, and 10, overlapped in t-SNE space, demonstrating that they would be smelled similarly by a rat. Compared with olfaction information maps, cluster overlapping is observed more in the molecular information maps (Fig. 6b). Interestingly, some clusters, such as clusters 1, 9, and 12, are clustered into multiple groups in the olfactory map, but are clustered into a single group in the molecular information map. For example, the olfactory images for odorants in cluster 12-1 are different from those in cluster 12-2 because of the distance between these two groups in the olfaction information map. However, in the molecular information map, odorants in cluster 12 are clustered into only one group. This demonstrates that 46 parameters are not enough for ideal mapping, and that some vital functional group parameters could have been missing from the analysis.

3.4. *Functional group-identification models*

To assess the potential for OM and MP to classify odorants, we applied LVQ, SVM, and ELM classification methods to each and compared the results. The features acquired by PCA or t-SNE were set as the input data for 178 odorants, and the 14 types of functional group labels were set as the output of the models.

3.4.1. *LVQ models*

To acquire optimized LVQ models, training epoch, learning rate, and learning goal were set to 500, 0.1, and 0.05, respectively. As a necessary parameter for LVQ networks, the number of hidden nodes is always set by trial and error. Here, the range of hidden layer nodes was 1–100 (Fig. S3). We chose the smallest numbers because these were associated with greater accuracy. The numbers of hidden layer nodes for OM-PCA, OM-tSNE, MP-PCA, and MP-tSNE were 35, 15, 34, and 98, respectively (Table 2).

The accuracy of functional group identification for calibration and validation sets using the four types of datasets are listed in Table 3. The results suggest that models calibrated by PCA higher

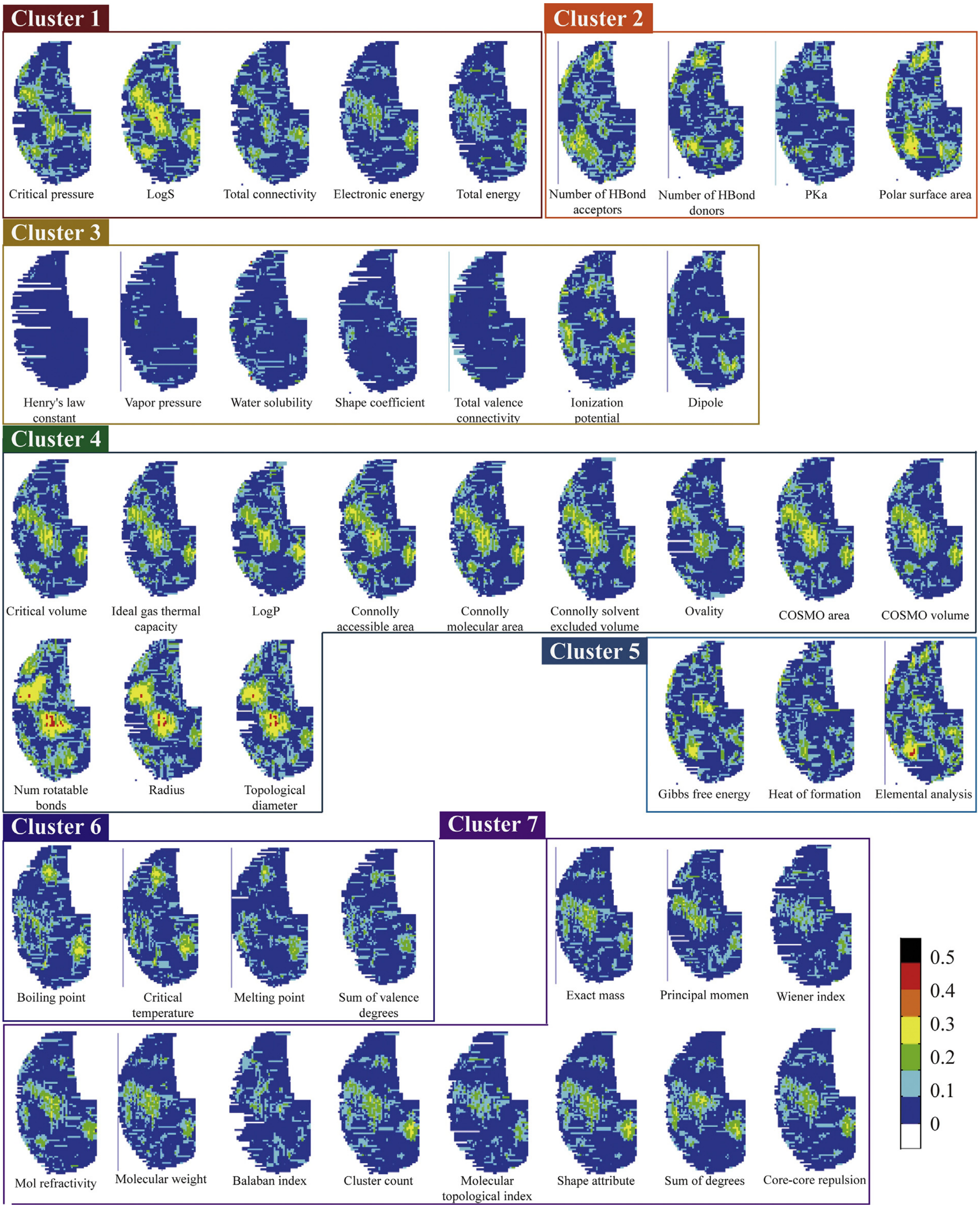

Fig. 5. The PCC-maps for 46 molecular parameters. The value on each pixel of a PCC-map indicates the correlation coefficient between a pixel and a molecular parameter. All PCC-maps are clustered seven groups.

Table 2
The optimal training parameters of LVQ, SVM and ELM models.

Input data		LVQ			SVM		ELM		
Odorant descriptor	Pretreatment methods	Input layer nodes	Hidden layer nodes	Output layer nodes	*c*	*g*	Input layer nodes	Hidden layer nodes	Output layer nodes
Odor maps	PCA	80	35	14	1.000	0.100	80	54	14
	t-SNE	2	15	14	1.000	0.100	2	51	14
Molecular parameters	PCA	23	34	14	5.657	0.354	23	21	14
	t-SNE	2	98	14	22.627	11.314	2	47	14

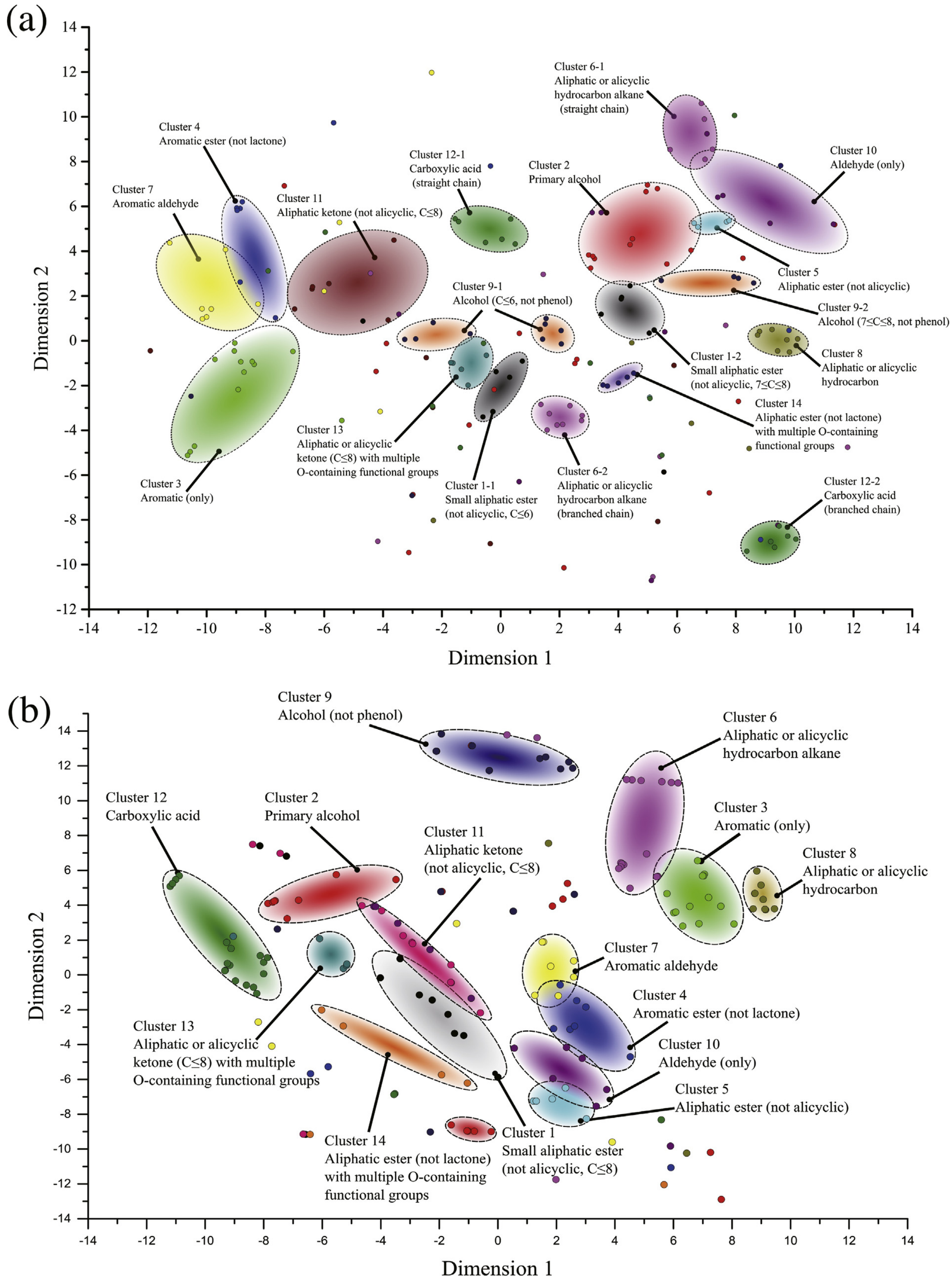

Fig. 6. Odorant clustering generated in OM (a) and MP (b) spaces by using t-SNE method. Dimension 1 and 2 indicates 2 values calculated by t-SNE. Each point indicates an odorant.

Table 3
Identification accuracies of LVQ, SVM and ELM models.

Odorant descriptor	Pretreatment methods	Modeling approach	Accuracy (%)		
			Calibration set	Validation set	Mean
Odor maps	PCA	LVQ	72.09	68.89	70.49
		SVM	29.63	18.60	24.12
		ELM	94.81	93.02	93.92
	t-SNE	LVQ	14.81	13.95	14.38
		SVM	14.81	13.95	14.38
		ELM	72.59	90.70	81.65
Molecular parameters	PCA	LVQ	65.93	69.77	67.85
		SVM	93.33	90.70	92.02
		ELM	89.63	93.02	91.33
	t-SNE	LVQ	21.48	23.26	22.37
		SVM	82.22	83.72	82.97
		ELM	86.67	95.35	91.01

accuracies than those calibrated by t-SNE. Compared with t-SNE, LVQ thus more suitable for the dataset obtained by PCA. However, LVQ identification of odorants generally poor, with the highest accuracy being only 69.77%.

3.4.2. SVM models

The SVM kernel function was set to RBF, and 5-fold cross validation was performed to obtain the penalty factor (c) and the RBF parameter (g). The parameters for the SVM models are listed in Table 2, and the accuracy of each SVM model is shown in Table 3. For the calibration set, the accuracies for the MP-PCA-SVM and MP-tSNE-SVM models are 93.33% and 82.63%, respectively, which higher than those observed for the OM-PCA-SVM and OM-tSNE-SVM models (29.63% and 14.81%, respectively). Similar results are shown for the validation sets. This demonstrates that the SVM models calibrated by molecular information performed better than those calibrated by olfactory information.

3.4.3. ELM models

We set the excitation function for the ELM models to "sig." The number of hidden layer nodes for ELM models were also determined by trial and error. Here, one model was trained 1000 times to overcome the randomness of ELM models. The numbers of hidden layers were chosen based on the average accuracy across the 1000 models (Fig. S4), and equaled 54, 51, 21 and 47 for the OM-PCA, OM-tSNE, MP-PCA, and MP-tSNE models respectively (Table 2). The accuracy of functional group identification for calibration and validation sets using different datasets is shown in Table 3. The results show that for the calibration set, the OM-PCA-ELM model was more accuracy (94.81%) than the other models (OM-tSNE-ELM, 72.59%; MP-PCA-ELM, 89.63%; MP-tSNE-ELM, 86.67%). For the validation set, all models were more than 90% accuracy. The MP-tSNE-ELM model was the most accuracy (95.35%), followed by the OM-PCA-ELM and MP-PCA-ELM models (93.02%), and the OM-tSNE-ELM model (90.70%). Thus, the OI-tSNE-ELM, MP-PCA-ELM, and MP-tSNE-ELM models were more accuracy for the validation set than for the calibration set. Although validation-set accuracy is generally lower than calibration-set accuracy, the reverse is possible if most of the represented samples were chosen in the calibration set that established the model [58,59].

3.4.4. Identification performance for different models

Comparing the three types of models, ELM was the best at functional group identification. This was likely because ELM does well in generalization. This is consistent with other studies showing good prediction performance by ELM [60,61]. We also found that SVM performed well with MP-PCA (93.33% and 90.70%) and MP-tSNE (82.22% and 83.72%) datasets. However, poor results were observed for OM-PCA-SVM (29.63% and 18.60%) and OM-tSNE-SVM (14.81% and 13.95%) models. This suggests that SVM models are more suitable for establishing functional group-identification models based on molecular parameters. Although MP-tSNE-ELM had the highest accuracy (95.35%) for the validation set, its accuracy in the calibration set was only 86.67%. The model calibrated by OM-PCA-ELM presented acceptable identification accuracies for both the calibration (94.81%) and validation (93.02%) sets. Therefore, we suggest that OM-PCA-ELM is the optimal model for identifying functional groups of odorants. Compared with other datasets, for functional group identification in odorants, the features extracted from odor maps via PCA contained more information than the 46 molecular parameters.

3.5. Discussion

An odorant can be described by multiple molecular parameters or by a neuronal response pattern in the mammal OB. Investigating the relationship between molecular features and OB-derived images of neuronal activity is a challenge in developing sensor-based machine olfaction. Many analyses have been carried out focusing on the classification of odor descriptors or molecular feature [62–64]. The importance of the olfactory information in OB is not fully understood and thus less attention has rarely been paid to the image analysisi of OB. In our previous study that was based on PCA, 15 key parameters were obtained by evaluating the correlations between the molecular parameters and PCs [53]. However, only six PCs were analyzed which might not be enough for describing all odorant features due to the complexity and nonlinearity of the dataset. In the present study, t-SEN was applied for the high-dimensional data analysis. t-SNE was considered to be able to provide a competitive performance in dimensionality reduction if compared with conventional methods such as PCA and multidimensional scaling. The result shown in this study confirm that t-SNE can be used as an effective approach to establish relationship between molecular parameters, odor map and functional groups.

The odor clustering in olfactory bulb is not well understood due to the limited amount of reported odorant molecules. To discover all the primary clusters may be a massive undertaking for clarify the mechanism of olfactory perception. The more data that can be acquired, the higher the model accuracy will be. In this study, only 178 odorants with non-intersecting labels were considered. Therefore, the model is applicable for limited types of chemicals. The current model cannot predict the functional groups for molecules with high molecular weight and complicated structures. In the future, more odor-response images of neuronal activity in the OB will be investigated to determine the hidden patterns in t-SNE space. Additionally, a larger variety of molecular parameters will be considered so that the possibility of describing an odorant by molecular information closer to the mammal olfaction.

4. Conclusions

In this study, we accumulated and analyzed 178 odor maps from the LJ database and their 46 types of molecular features. PCC-maps for molecular parameters turned out to be clustered in seven groups, and the parameters in each group had a similar effect on the images of olfactory responses. Low correlation coefficients indicated that the relationship between molecular features and the odor map responses was not linear. All odorants were mapped in 2D space, and similar odorants were clustered together. Compared with the cluster map generated by molecular parameters, olfactory images contained more detailed information, such as the lengths of carbon and branched chains. We tested how well different models could identify functional groups when the models were calibrated based on olfactory information or molecular parameters. The results showed that OM-PCA-ELM was the optimal model. Although models calibrated by molecular parameters were weaker than those based on odor maps, a comparative model could be established if it was based on enough molecular features. This research supports the structure-odor relationship from a data-analysis perspective.

Acknowledgments

This research was supported by a grant from China Scholarship Council (CSC) and JSPS KAKENHI Grant Number 15H01713.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2017.08.024.

References

[1] K. Mori, Y.K. Takahashi, K.M. Igarashi, M. Yamaguchi, Maps of odorant molecular features in the mammalian olfactory bulb, Physiol. Rev. 86 (2006) 409–433.
[2] C. Bushdid, M.O. Magnasco, L.B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli, Science 343 (2014) 1370–1372.
[3] S.X. Luo, R. Axel, L.F. Abbott, Generating sparse and selective third-order responses in the olfactory system of the fly, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 10713–10718.
[4] L.B. Buck, Information coding in the vertebrate olfactory system, Annu. Rev. Neurosci. 19 (1996) 517–544.
[5] R. Vassar, J. Ngai, R. Axel, Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium, Cell 74 (1993) 309–318.
[6] M. Stopfer, S. Bhagavan, B.H. Smith, G. Laurent, Impaired odour discrimination on desynchronization of odour-encoding neural assemblies, Nature 390 (1997) 70–74.
[7] J.A. Gottfried, J. O'Doherty, R.J. Dolan, Encoding predictive reward value in human amygdala and orbitofrontal cortex, Science 301 (2003) 1104–1107.
[8] B.D. Rubin, L.C. Katz, Optical imaging of odorant representations in the mammalian olfactory bulb, Neuron 23 (1999) 499–511.
[9] M. Wachowiak, L.B. Cohen, Representation of odorants by receptor neuron input to the mouse olfactory bulb, Neuron 32 (2001) 723–735.
[10] E.R. Soucy, D.F. Albeanu, A.L. Fantana, V.N. Murthy, M. Meister, Precision and diversity in an odor map on the olfactory bulb, Nat. Neurosci. 12 (2009) 210–220.
[11] P. Mombaerts, F. Wang, C. Dulac, S.K. Chao, A. Nemes, M. Mendelsohn, et al., Visualizing an olfactory sensory map, Cell 87 (1996) 675–686.
[12] N. Uchida, Y.K. Takahashi, M. Tanifuji, K. Mori, Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features, Nat. Neurosci. 3 (2000) 1035–1043.
[13] K. Snitz, A. Yablonka, T. Weiss, I. Frumin, R.M. Khan, N. Sobel, Predicting odor perceptual similarity from odor structure, PLoS Comput. Biol. 9 (2013) e1003184.
[14] B. Malnic, J. Hirono, T. Sato, L.B. Buck, Combinatorial receptor codes for odors, Cell 96 (1999) 713–723.
[15] B.A. Johnson, H. Farahbod, M. Leon, Interactions between odorant functional group and hydrocarbon structure influence activity in glomerular response modules in the rat olfactory bulb, J. Comp. Neurol. 483 (2005) 205–216.
[16] B.A. Johnson, H. Farahbod, S. Saber, M. Leon, Effects of functional group position on spatial representations of aliphatic odorants in the rat olfactory bulb, J. Comp. Neurol. 483 (2005) 192–204.
[17] K. Mori, H. Nagao, Y. Yoshihara, The olfactory bulb: coding and processing of odor molecule information, Science 286 (1999) 711–715.
[18] K. Kaeppler, F. Mueller, Odor classification: a review of factors influencing perception-based odor arrangements, Chem. Senses (2013), bjs141.
[19] R. Pellegrino, P.G. Crandall, H.S. Seo, Using olfaction and unpleasant reminders to reduce the intention-behavior gap in hand washing, Sci. Rep. 6 (2016).
[20] M. Wienisch, V.N. Murthy, Population imaging at subcellular resolution supports specific and local inhibition by granule cells in the olfactory bulb, Sci. Rep. 6 (2016).
[21] M. Zarzo, D.T. Stanton, Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps, Attention Percept. Psychophys. 71 (2009) 225–247.
[22] M. Korichi, V. Gerbaud, T. Talou, P. Floquet, A.H. Meniai, S. Nacef, Computer-aided aroma design. Ii. Quantitative structure-odour relationship, Chem. Eng. Process. 47 (2008) 1912–1925.
[23] R. Haddad, R. Khan, Y.K. Takahashi, K. Mori, D. Harel, N. Sobel, A metric for odorant comparison, Nat. Methods 5 (2008) 425–429.
[24] L. Secundo, K. Snitz, K. Weissler, L. Pinchover, Y. Shoenfeld, R. Loewenthal, et al., Individual olfactory perception reveals meaningful nonolfactory genetic information, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 8750–8755.
[25] B. Raman, R. Gutierrez-Osuna, Relating sensor responses of odorants to their organoleptic properties by means of a biologically-inspired model of receptor neuron convergence onto olfactory bulb, in: A. Gutiérrez, S. Marco (Eds.), Biologically Inspired Signal Processing for Chemical Sensing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 93–108.
[26] Y. Nozaki, T. Nakamoto, Odor impression prediction from mass spectra, PLoS One 11 (2016) e0157030.
[27] B. Auffarth, Understanding smell-the olfactory stimulus problem, Neurosci. Biobehav. Rev. 37 (2013) 1667–1679.
[28] K. Kaeppler, F. Mueller, Odor classification: a review of factors influencing perception-based odor arrangements, Chem. Senses 38 (2013) 189–209.
[29] C.A. Levitan, J. Ren, A.T. Woods, S. Boesveldt, J.S. Chan, K.J. McKenzie, et al., Cross-cultural color-odor associations, PLoS One 9 (2014) e101651.
[30] M. Falasconi, A. Gutierrez-Galvez, M. Leon, B.A. Johnson, S. Marco, Cluster analysis of rat olfactory bulb responses to diverse odorants, Chem. Senses 37 (2012) 639–653.
[31] B.A. Johnson, Z. Xu, S.S. Ali, M. Leon, Spatial representations of odorants in olfactory bulbs of rats and mice: similarities and differences in chemotopic organization, J. Comp. Neurol. 514 (2009) 658–673.
[32] W. Guo, J. Gu, D. Liu, L. Shang, Peach variety identification using near-infrared diffuse reflectance spectroscopy, Comput. Electron. Agric. 123 (2016) 297–303.
[33] X. Zhu, L. Fang, J. Gu, W. Guo, Feasibility investigation on determining soluble solids content of peaches using dielectric spectra, Food Anal. Method 9 (2016) 1789–1798.
[34] L. Van der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.
[35] J. Cheng, H.J. Liu, F. Wang, H.S. Li, C. Zhu, Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning, IEEE Trans. Image Process. 24 (2015).
[36] X.H. Li, W. Kong, W.M. Shi, Q. Shen, A combination of chemometrics methods and GC–MS for the classification of edible vegetable oils, Chemometrics Intellig. Lab. Syst. 155 (2016) 145–150.
[37] H. Kaneko, K. Funatsu, Preparation of comprehensive data from huge data sets for predictive soft sensors, Chemom. Intell. Lab. Syst. 153 (2016) 75–81.
[38] W. Gani, M. Limam, A kernel distance-based representative subset selection method, J. Stat. Comput. Sim. 86 (2016) 135–148.
[39] G.B. da Costa, D.D.S. Fernandes, A.A. Gomes, V.E. de Almeida, G. Veras, Using near infrared spectroscopy to classify soybean oil according to expiration date, Food Chem. 196 (2016) 539–543.
[40] T.H. Sun, F.C. Tien, F.C. Tien, R.J. Kuo, Automated thermal fuse inspection using machine vision and artificial neural networks, J. Intell. Manuf. 27 (2016) 639–651.
[41] T. Liu, C.S. Chen, X.Z. Shi, C.Y. Liu, Evaluation of Raman spectra of human brain tumor tissue using the learning vector quantization neural network, Laser Phys. 26 (2016).
[42] A. Bohnsack, K. Domaschke, M. Kaden, M. Lange, T. Villmann, Learning matrix quantization and relevance learning based on schatten-p-norms, Neurocomputing 192 (2016) 104–114.
[43] J. Cho, R. Anandakathir, A. Kumar, J. Kumar, P.U. Kurup, Sensitive and fast recognition of explosives using fluorescent polymer sensors and pattern recognition analysis, Sens. Actuators B 160 (2011) 1237–1243.
[44] R. Kumar, A.P. Bhondekar, R. Kaur, S. Vig, A. Sharma, P. Kapur, A simple electronic tongue, Sens. Actuators B 171 (2012) 1046–1053.
[45] Y.S. Wang, M. Yang, G. Wei, R.F. Hu, Z.Y. Luo, G. Li, Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy, Sens. Actuators B 193 (2014) 723–729.
[46] W.C. Guo, L. Shang, X.H. Zhu, S.O. Nelson, Nondestructive detection of soluble solids content of apples from dielectric spectra with ANN and chemometric methods, Food Bioprocess Technol. 8 (2015) 1126–1138.
[47] L. Zhang, F.C. Tian, H. Nie, L.J. Dang, G.R. Li, Q. Ye, et al., Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine, Sens. Actuators B 174 (2012) 114–125.
[48] C.C. Chang, C.J. Lin, Libsvmm: a library for support vector machines, ACM Trans. Intel. Syst. Technol. 2 (2011).

[49] G.B. Huang, H.M. Zhou, X.J. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. B 42 (2012) 513–529.
[50] G.B. Huang, D.H. Wang, Y. Lan, Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern. 2 (2011) 107–122.
[51] L. Cornejo-Bueno, J.C. Nieto-Borge, P. Garcia-Diaz, G. Rodriguez, S. Salcedo-Sanz, Significant wave height and energy flux prediction for marine energy applications: a grouping genetic algorithm – extreme learning machine approach, Renew. Energy 97 (2016) 380–389.
[52] L. Shang, W.C. Guo, S.O. Nelson, Apple variety identification based on dielectric spectra and chemometric methods, Food Anal. Method 8 (2015) 1042–1052.
[53] M. Imahashi, K. Hayashi, Odor clustering based on molecular parameter for odor sensing, Sens. Mater. 26 (2014) 171–180.
[54] Y.K. Takahashi, M. Kurosaki, S. Hirono, K. Mori, Topographic representation of odorant molecular features in the rat olfactory bulb, J. Neurophysiol. 92 (2004) 2413–2427.
[55] J.A. Gottfried, J.S. Winston, R.J. Dolan, Dissociable codes of odor quality and odorant structure in human piriform cortex, Neuron 49 (2006) 467–479.
[56] B.N. Johnson, J.D. Mainland, N. Sobel, Rapid olfactory processing implicates subcortical control of an olfactomotor system, J. Neurophysiol. 90 (2003) 1084–1094.
[57] P. Comon, Independent component analysis, a new concept, Signal Process. 36 (1994) 287–314.
[58] L. Turgeman, J.H. May, R. Sciulli, Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission, Expert Syst. Appl. 78 (2017) 376–385.
[59] S. Suresh, R.V. Babu, H.J. Kim, No-reference image quality assessment using modified extreme learning machine classifier, Appl. Soft Comput. 9 (2009) 541–552.
[60] M. Xia, Y.C. Zhang, L.G. Weng, X.L. Ye, Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs, Knowl. Based Syst. 36 (2012) 253–259.
[61] S.J. Lin, C.H. Chang, M.F. Hsu, Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction, Knowl. Based Syst. 39 (2013) 214–223.
[62] R. Kumar, R. Kaur, B. Auffarth, A.P. Bhondekar, Understanding the odour spaces: a step towards solving olfactory stimulus-percept problem, PLoS One 10 (2015) e0141263.
[63] F. Kermen, A. Chakirian, C. Sezille, P. Joussain, G. Le Goff, A. Ziessel, et al., Molecular complexity determines the number of olfactory notes and the pleasantness of smells, Sci. Rep. 1 (2011).
[64] D. Zakarya, D. Cherqaoui, M. Esseffar, D. Villemin, J.M. Cense, Application of neural networks to structure sandalwood odour relationships, J. Phys. Org. Chem. 10 (1997) 612–622.

Biographies

Liang Shang received the B.S. degree in mechanical manufacturing and automation and the M.S. degree in agricultural mechanization engineering from Northwest A&F University, Shannxi, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree at Kyushu University, Fukuoka, Japan, and engaging in research related to odor sensors and machine learning.

Chuanjun Liu received his PhD degree in material engineering from Nagaoka University of Technology (Japan) in 2006. He has worked as research fellow in Nagaoka University of Technology (from 2006) and Kyushu University (from 2008), and as assistant professor at the Graduate School of Information Science and Electrical Engineering of Kyushu University (from 2012) and associate professor of R&D center for Taste and Odor Sensing (TAOS) of Kyushu University (from 2016). He is now a principle researcher in Research Laboratory of U.S.E. Co. LTD, Japan. He is a member of the Society of Polymer Science Japan and the Institute of Electrical Engineers of Japan. His research interests include the development and application of organic electronic devices, nanoscale sensing materials, and gas and odor sensors.

Yoichi Tomiura is a professor at the Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University. His interests include: disambiguation of syntactic structure of a sentence, acquisition of the knowledge about words' meaning from a corpus, organization of documents on WWW, and a use of documents on WWW to the language education.

Kenshi Hayashi received his BE, ME, and PhD degree, all in electrical engineering, from Kyushu University (Japan) in 1982, 1984, and 1990, respectively. He is now a professor at the Graduate School of Information Science and Electrical Engineering of Kyushu University. He is a member of the Japan Society of Applied Physics and the Institute of Electrical Engineers of Japan. His research interests include: information service using odor cluster matching, visualization of odor space, measurement and coding of odor quality and quantity, novel devices using molecular wire and organic electronic material, and biometrics by odor sensing.