Machine-Learning-Based Olfactometry: Odor Descriptor Clustering Analysis for Olfactory Perception Prediction of Odorant Molecules

Liang Shang,* Chuanjun Liu, Fengzhen Tang,* Bin Chen, Lianqing Liu, Kenshi Hayashi

Abstract: Although gas chromatography/olfactometry (GC/O) has been employed as a powerful analytical tool in odor measurement, its application is limited by the variability, subjectivity, and high cost of the trained panelists who are used as detectors in the system. The advancements in data-driven science have made it possible to predict structure-odor-relationship (SOR) and thus to develop machine-learning-based olfactometry (ML-GCO) in which the human panelists may be replaced by machine learning models to obtain the sensory information of GC-separated chemical compounds. However, one challenge that remained in ML-GCO is that there are too many odor descriptors (ODs) being used to describe the sensory characteristics of odorants. It is impractical to build a corresponding model for each OD. To solve this issue, we propose a SOR prediction approach based on odor descriptor clustering. 265 representative ODs are firstly classified into 20 categories using a co-occurrence Bayesian embedding model. The categorization effect is explained according to the semantic relationships using a pre-trained Word2Vec model. Various molecular structure features including molecularly parameters, molecular fingerprints, and molecular 2D graphic features extracted by convolutional neural networks, are employed to predict the aforementioned odor categories. High prediction accuracies (Area under ROC curve was 0.800±0.004) demonstrate the rationality of the proposed clustering scenario and molecular feature extraction. This study makes the ML-GCO models much closer to the practical application since they can be expected as either an auxiliary system or complete replacement of human panelists to perform the olfactory evaluation.

Gas chromatography/olfactometry (GC/O) is a key technique that integrates the separation of volatile compounds using a gas chromatograph (GC) with the detection of odor employing human assessors as olfactometers[1]. Contributed by mass spectrometry (MS) analysis, GC/O can provide not only the molecular information of complex odor mixtures, but also sensory information of specific odor-active components. Therefore, it has been applied as a critical analytical instrument in various filed, such as food, cosmetics, agriculture, and environment[2].

In GC/O analysis, human assessors play a decisive role. It has, however, been indicated that the major problems of olfactometry are variability, subjectivity, and the high cost of training and employing human panelists. In general, the GC/O measurement needs an odor evaluation team containing 4~8 assessors who need to remember an odor panel composed of 6~8 pivotal odor descriptors (ODs) selected by specialists to calibrate their odor perception memory. They are requested to make sensory evaluations through sniffing GC effluent components, respectively. Finally, the sensory evaluation results are summarized and analyzed. Because of the lengthy sample preparations and MS analysis, panelists are usually on standby for the predecessor task finished. Moreover, the subjectivity of panelists at the intra- and inter-individual level is also an inevitable problem of existing GC/O. Recently, it has been reported that computational approaches developed as an assistance system for human assessors would reduce the above problems[3].

The precise relationship between molecular structure and odor perception, termed the structure-odor relationship (SOR), has attracted considerable attention from the aspect of data science[4]. Great efforts have been made in ODs identification based on various molecular features, such as physicochemical parameters, bio-inspired olfactory model, and MS[5]. Moreover, various deep neural network models (DNNs), machine learning (ML) classification frameworks, and topic models have been developed to express the SOR[6]. The abovementioned studies indicated that the SOR would be solved by ML algorithms and

chemometrics. Therefore, our research team previously developed a method for conducting ML-based GC/O (ML-GCO), in which olfactometry detection can be done by an ML classifier, thus reducing the dependence on a human panelist[7].

It should be mentioned that, however, there are many problems needed to be overcome before the practical application of ML-GCO. For example, the dimension of odor space remains unknown and it is not yet clear on the primary dimensions of olfactory like vision or gustatory[8]. Odors are highly complex and people are known to disagree regarding their linguistic descriptions of smell sensations. The number of odor compounds is estimated at over 400,000 which are described by several hundred odor descriptors[9]. In most reported work, only a small number of typical, common odor descriptors have been included, leaving the majority of them undefined [7, 10, 11]. From a practical point of view, it is impressible to build the corresponding prediction models for so many individual descriptors. Many researches have demonstrated that odor descriptors are inherently connected in the perceptual space[12]. For example, Kumar et al., have developed a graphic method to draw an odor network based on the similarity between odor descriptors[11]. Villière et. al. propose the SketchOscent, a hierarchical visual and interactive representation of the odorous space derived from a knowledge-based model[13]. These results demonstrate the odor descriptors can be clustered according to their similarity evaluations and thus cover the odor descriptor space as many as possible. The SOR prediction based on the odor descriptor clustering analysis may be a good way to solve the practicality of ML-GCO.

In response to the above problems, ODs embedding and clustering approaches are proposed in this study. A schematic of the data-processing method is illustrated in Figure 1. ODs from different data sources are collected and merged, and Bayesian co-occurrence embedding (BCE) was employed for odorants and ODs embedding. Based on the distance between odorant vectors and OD vectors, ODs can be calibrated and relabeled. And then hierarchical clustering analysis (HCA) was performed on the embedded OD vectors to investigate the internal relationship between ODs and their clustering result is discussed. We show that a total of 265 ODs were clustered in 20 categories, and most of the categories can be supported by previous research. In addition, the structure parameters of odorant molecules are employed for training ML models to verify the rationality of the proposed clustering scenario. Results show that the smell categories can be predicted by the models established by molecularly structure features of odorants successfully (Area under ROC curve: 0.800±0.004, precision: 0.595±0.004, recall: 0.721±0.003 and F-score: 0.570±0.007, $p<0.0001$ Wilcoxon test). It indicated that the smell clusters proposed in this study are supported by the SOR. The proposed odor category is not only expected as a novel research direction for developing ML-GCO, but also applies a reference for understanding the biology of olfaction.

■ MATERIALS AND METHODS

Data Preparation. To understand the internal relationships between smell percepts, we collected the chemical abstracts service (CAS) data of odorants and their odor perceptions via both web scraping and manual methods. We used three publically available odor databases, including the Flavors and Fragrances database (Sigma-Aldrich)[14], the Good Scents database[15], and the Odor Map database[16]. Detailed information of odor databases used in the present study is summarized in Table 1. Based on CAS number, the simplified molecular-input line-entry system (SMILES) data were collected from PubChem (https://pubchem.ncbi.nlm.nih.gov)[17]. Using RDKit software (http://www.rdkit.org)[18], molecular structure images, molecular parameters, and molecular fingerprints were obtained for further analysis. In the present study, chemicals without odor descriptors or those considered "odorless" were not considered. Finally, 2849 molecules with 510 ODs were collected and analyzed. All of the ODs in the database are listed in Table S1.

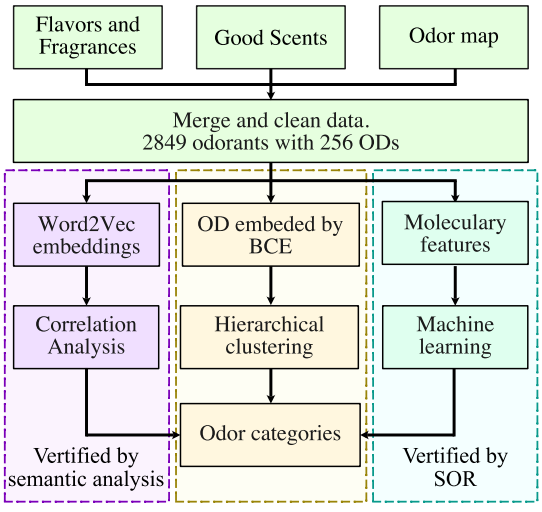

Figure 1. Data processing diagram of odor descriptor clustering analysis.

Table 1. Summary of Odorant Databases Used in Present Study.

data base	# odorants	description
Flavors and Fragrances	1026	An odor database of flavor and fragrance proposed by Sigma Aldrich.
Good Scents	3673	An odor database of flavor, fragrance, food and cosmetic industries.
OdorMapDB	321	An odor database of odorants and their olfactory bulb responses (odor maps).

Bayesian Co-occurrence Embedding. To normalize the labels for odorants from all databases, BCE was introduced for OD vector embedding[19]. A brief description of BCE is illustrated in Figure 2. Based on BCE, a preference matrix for each odorant can be generated. A plus sign (+) indicates comparisons with OD_d, OD_j is the label for the odorant$_i$, a minus sign (-) indicates the opposite, and a question mark (?) indicates the unknown value that is estimated by the BCE algorithm. Therefore, the individual probability that an odorant prefers OD_d to OD_j was defined as:

$$P(d >_C j|\Theta) := \sigma(\hat{x}_{Cdj}(\Theta)) \tag{1}$$

where σ is the logistic sigmoid function, Θ is the parameters of the BCE model, and $\hat{x}_{Cdj}(\Theta)$ is the score indicating the degree to which chemical C prefers OD_d rather than OD_j. According to maximum posterior estimation, the generic optimization criterion for each odorant could be estimated as follows:

$$\begin{aligned}\text{BCE Loss} : &= \ln p(\Theta|>_c) \\ &= \ln p(>_c|\Theta)P(\Theta) \\ &= \ln \prod_{(c,d,j)\in D_s} \sigma(\hat{x}_{cdj})P(\Theta) \\ &= \sum_{(c,d,j)\in D_s} \ln \sigma(\hat{x}_{cdj}) + \ln P(\Theta) \\ &= \sum_{(c,d,j)\in D_s} \ln \sigma(\hat{x}_{cdj}) - \lambda_\Theta \|\Theta\|^2\end{aligned} \tag{2}$$

where λ_Θ are model-specific regularization parameters. Therefore, the gradient of the loss function with respect to the model parameters is:

$$\begin{aligned}\frac{\partial \text{BCE Loss}}{\partial \Theta} &= \sum_{(c,d,j)\in D_s} \frac{\partial}{\partial \Theta} \ln \sigma(\hat{x}_{cdj}) - \lambda_\Theta \frac{\partial}{\partial \Theta} \|\Theta\|^2 \\ &\propto \sum_{(c,d,j)\in D_s} \frac{-\exp(-\hat{x}_{cdj})}{1+\exp(-\hat{x}_{cdj})} \cdot \frac{\partial}{\partial \Theta} \hat{x}_{cdj} - \lambda_\Theta \Theta\end{aligned} \tag{3}$$

Finally, the model parameters (Θ) could be updated using the assigned learning rate η and stochastic gradient descent as follows.

$$\Theta \leftarrow \Theta - \eta\left(\frac{-\exp(\hat{x}_{cdj})}{1+\exp(-\hat{x}_{cdj})} \cdot \frac{\partial}{\partial \Theta} \hat{x}_{cdj} - \lambda_\Theta \Theta\right) \tag{4}$$

Unlike k-nearest neighbor (kNN) collaborative filtering or matrix factorization (MF), BCE applies a Bayesian optimization criterion to generate odorant similarity rankings based on pairs of ODs (i.e. the odorant-specific order of two ODs). As an offline embedding method, Bayesian optimization has advantages over the standard learning techniques for MF and kNN[20]. In the present study, the relabeling results produced by BCE were evaluated via the normalized discounted cumulative gain (NDCG).

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{5}$$

where the discounted cumulative gain (DCG) and ideal discounted cumulative gain (IDCG) were calculated as:

$$\begin{aligned}DCG_p &= \sum_{i=1}^{p} \frac{2^{rel_i}-1}{\log_2(i+1)} \\ IDCG_p &= \sum_{i=1}^{REL_p} \frac{2^{rel_i}-1}{\log_2(i+1)}\end{aligned} \tag{6}$$

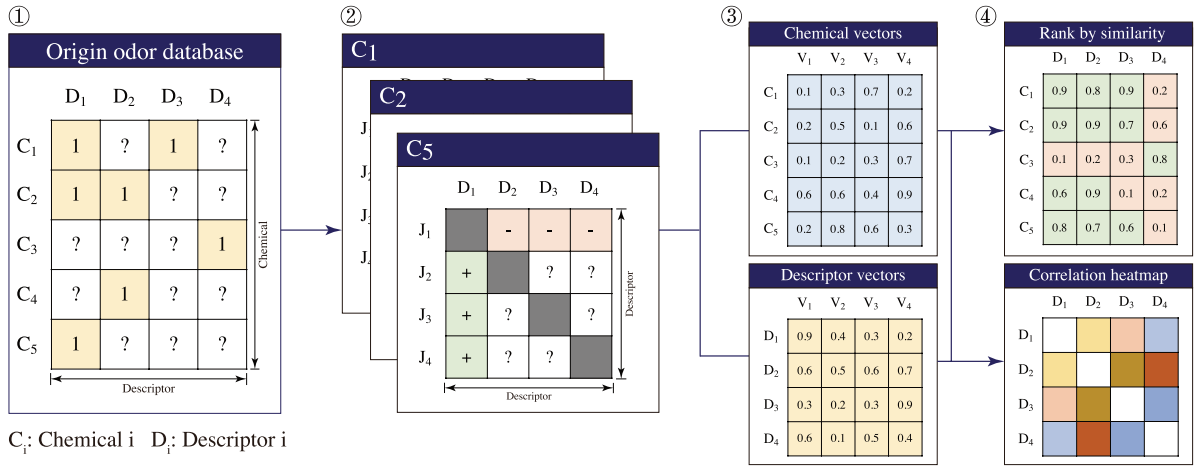

Figure 2. Odor descriptor embedding and calibration using the Bayesian co-occurrence embedding (BCE) method.

Molecular Graphic Feature Extraction. To extract the necessary numerical features from the molecular structure images, we considered four types of pretrained convolutional neural network (CNN) frameworks, including VGG-16, Restnet, Densnet, and Alexnet. Detailed information for these CNNs is given in these papers[21]. In the present study, the CNNs were only applied as molecular-graphic feature extractors for the SOR model calibration.

SOR Model Calibration. The dataset for the odor category prediction was a typical imbalanced data set because the class distribution of the positive samples (minor samples with specified odor categories) and negative samples (major samples with non-specified odor categories) was not uniform. Inspired by Mordelet's research, we considered bagging classifiers to be a feasible method for learning with an imbalanced data set[22]. Detail description for transductive bagging learning is presented in the support information. In the present study, the sample pool was divided into training and test sets with a 3:1 ratio via random stratified sampling. In addition, the number of bootstraps was set as 100 and the subsample number was the same as that for the positive numbers. Considering the sample size, we employed GBDT and GLVQ to predict the odor clusters in the present study[23]. More introduction for these models is described in the support information. Finally, the optimal feature extractor and model combination was determined by considering the area under the ROC curve (AUC-ROC), precision, recall, and F1-score of the test set, respectively. Detailed information for these metrics is presented in the support information materials. In the present study, data and models were processed and analyzed using Python (ver. 3.9.0) and R (ver. 4.1.1).

Word2Vec Embedding Model. To investigate the semantic internal relationships between ODs for each odor category, we used Google's pre-trained Word2Vec model to create semantic presentations for ODs. The model, which contains 300-dimensional vectors for 3 million words and phrases, was trained using the Google News dataset. Two types of model architectures, including the continuous bag-of-words (CBOW) model and the continuous skip-gram model, were developed for learning latent presentations for words. More detailed information can be found here[24]. The proposed model has been confirmed to perform better than previous techniques based on different types of neural networks. In addition, the proposed vectors provide the state-of-the-art performance for measuring semantic word similarities.

■ RESULTS AND DISCUSSION

Odor Perception Embedding and Calibration. In the present study, the idea of co-occurrence in BCE was introduced for odor perception calibration. Using the BCE method, ODs can be embedded as numerical vectors. According to the cosine similarity between the odorant and OD vectors, the top 20 nearest ODs were considered as candidates for each odorant. As a critical factor, optimization of the embedded dimension is an important consideration. The normalized discounted cumulative gain for the top 20 ODs (NDCG@20) under different embedded dimensions is shown in Figure S1. The NDCG@20 increased with the number of embedded dimensions. Embedded vectors with 64 dimensions performed significantly better than embedded vectors with 32, 16, and 8 dimensions ($p<0.001$), and were not significantly different from embedded vectors with 128, 256, 512 dimensions. Considering accuracy and computational efficiency, we selected 64 as the optimal number of embedded dimensions for the BCE in the present study. The distribution of ODs before and after calibration is shown in Figure S2, and detailed information is given in Table S1. In summary, the sample size of the ODs increased after calibration in 70.58 % of cases. In this type of statistics, the sample size should always be more than 20. Using the BCE algorithm, the sample size increased to more than 20 in 61 ODs (11.96 %), which were then considered for further analysis. Finally, 265 ODs (51.96 %) were selected for afterward clustering analysis in the present study.

Clustering Characterization for Odor Descriptors. To quantify the inner relationships between ODs, we performed hierarchical clustering based on the Euclidean distances between the embedded vectors of smell descriptors. Because the basis of olfaction has not yet been established, we turned to previous research to explain our clustering results. The well-known Dravnieks[25] and DREAM[26] datasets include 19 types of descriptors: the scent of a bakery, sweet, fruit, fish, garlic, spices, cold, sour, burnt, acid, warm, musky, sweaty, ammonia/urinous, decayed, wood, grass, floral, and chemical[27]. Based on an analysis of previous research[28], we considered 20 to be a reasonable number of clusters. The results were organized and depicted using dendrograms, as shown in Figure 3 and Figure S3, S4. In the clustering results, the descriptors

with semantic similarity were almost always grouped in the same class. Specifically, cluster-1 was composed of sweaty-like and fish-like descriptors, which were considered unpleasant odors. ODs related to the scent of a bakery were clustered in cluster-2, and burnt-like descriptors were present in cluster-4 and cluster-12. Most groups, including cluster-5 (milky-like), cluster-6 (spicy-like), cluster-7 (musky or green-like), cluster-10 (wine or ester-like), cluster-11 (fruit acid-like), cluster-13 (cold or fresh-like), cluster-14 (garlic-like), cluster-15 (fruit-like), cluster-16 (floral), cluster-17 (cold-like), cluster-18 (musky and wood-like), and cluster-20 (wood-like), were supported by the core smell descriptors proposed in previous research[29].

Although most of the descriptors with similar linguistic meanings were clustered in the same group, special cases were also observed in some clusters. For cluster-3, the ODs included fruity, sweet, floral, etc., and so we regarded this as the sweet or fruit-like group. However, some descriptors related to plants, such as spicy, woody, herbal, and green, were also present in cluster-3. In addition, we found tomato, chrysanthemum, rotten cabbage, and radish in cluster-14, and labeled this as the garlic-like group. To investigate the reasons for this "mis-clustering" of smell descriptors, the odorants contained in these ODs were extracted and analyzed. For example, the odor of methyl mercaptan (CAS: 74-93-1) is reminiscent of garlic or rotten cabbage, while Erucin (CAS: 4430-36-8) and Berteroin (CAS: 4430-42-6) are labeled as the odors of cabbage and radish. The link between "garlic" and "rotten cabbage", and between "rotten cabbage" and "radish" indicated that these three descriptors could be clustered together, which can explain the presence of cabbage and radish in garlic-like clusters. In addition, Methyl propyl disulfide (CAS: 2179-60-4) is labeled as "radish", "mustard", "tomato", "garlic", etc., and Ethyl methyl sulfide (CAS: 2179-60-4) is labeled as "garlic", "tomato", "rotten cabbage", etc. Thus, the odor semantic database showed an internal connection between the smell of tomato and that of garlic, which could be explained by the latent relationships between the ODs. The above-mentioned analyses demonstrated that semantic descriptor clustering based on embedded vectors using the BCE method is a reasonable approach for understanding internal affiliations. Even though most of the clusters could be explained by common sense or previous research, some clusters, such as cluster-8, were not found to belong to any 'well-known' smell categories proposed by previous studies. In cluster-8, descriptors such as tea, root, basil, thyme, and buchu could be regarded as woody or herbal. Furthermore, fruit-like (mango, blueberry, passion fruit) and mild descriptors (faint, bland, slightly) were also clustered here. The plants associated with these ODs, such as tea or basil, are mild and herbaceous, which could explain this clustering result. In cluster-10, most of the descriptors were related to alcohol, such as wine-like, powerful, rum-like, alcoholic, etc. Wine flavor descriptors, such as berry, juicy, jam, candy, ether, and ester, were also identified. Consequently, cluster-10 was considered to have the aroma of wine. Chemical-like descriptors, including oil, chemical, and fusel, were also clustered here, which indicated that cluster-10 comprised multiple types of smell perceptions. Interestingly, unpleasant smell descriptors, such as sweaty/fish-like (cluster-1) and garlic-like (cluster-14), could be easily clustered, which confirmed that flavor-like impressions were regarded as more complex than unpleasant perceptions.

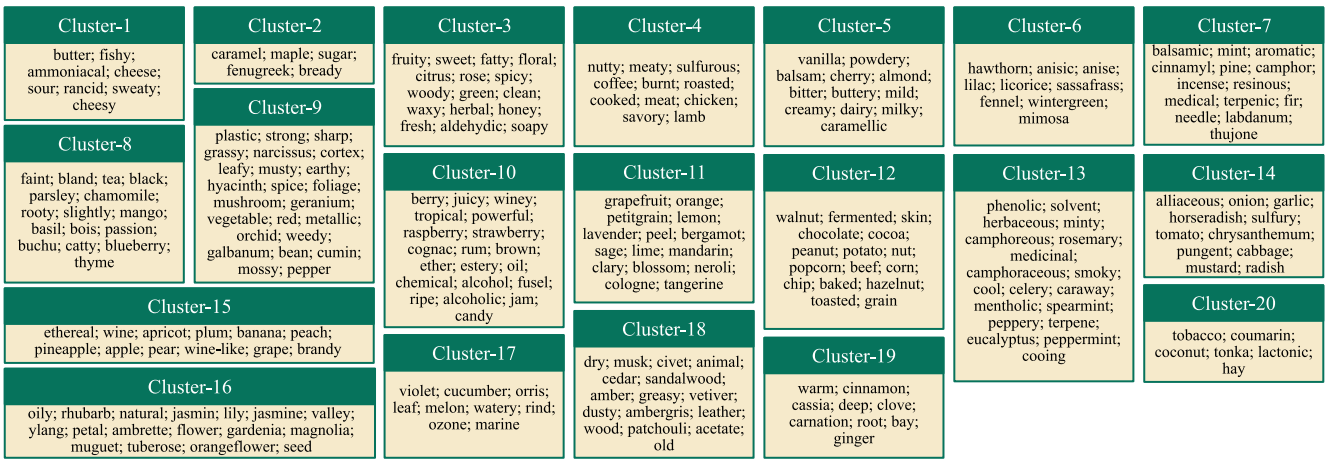

Figure 3. A summary of odor descriptor clustering based on the similarities between BCE embedded vectors.

Mapping Odor Descriptors in t-SNE Space. To investigate the internal relationships between odor categories, we visualized the data using Barnes-Hut t-distributed stochastic neighbor embedding (t-SNE) as an unsupervised low dimension presentation method. Detail explanation for the t-SNE method is described in support information. As illustrated in Figure 4, odor descriptors from 20 categories were mapped in t-SNE space via manifold embedding. The details for the embedded data calculated using t-SNE are presented in Table S2. The analysis illustrated that ODs from cluster-1 (sweaty or fish-like), cluster-2 (bakery-like), cluster-3 (woody or herb-like), cluster-7 (musky or green-like), cluster-11 (fruit or acid-like), cluster-12 (burnt-like), and cluster-14 (garlic-like) were clustered together. Additionally, descriptors from cluster-4 (burnt-like), cluster-13 (cold or fresh-like), cluster-16 (floral), and cluster-20 (wood-like) were clustered in multiple groups. For cluster-4, we found that part-1 (including chicken, lamb, and savory) was on the left of the t-SNE map, and part-2 (including nutty, meaty, sulfurous, coffee, burnt, roasted, cooked, and meat) was on the right of part-1. In addition, we found that beef was located near part-1 of cluster-4, which could be explained by the meat-like smell of beef. Smoky was located near part-2 of cluster-4, which was close to perceptions of burnt or roasted. In summary, plants or herb-related smell perception categories, such as cluster-6, cluster-7, cluster-8, cluster-13, and cluster-19, were located at the top area of the t-SNE space. In addition, fruity or alcohol-like categories (including cluster-10, cluster-11, and cluster-15) were neighbors, and were below the plant of herb-related categories. This demonstrates that the relationships between categories can be illustrated in t-SNE space, and that their locations can also be explained by their semantic meanings.

Comparison with Word2Vec Semantic Embeddings. Odor descriptors are not only used to express odor feelings, but are also applied in daily written communication. To explore the pure semantic relationships for each category, we conducted a linguistic analysis. We employed a pre-trained Word2Vec model provided by Google as a feature extractor to generate word vectors containing 3 million words based on roughly 100 billion words from a Google News dataset. Correlation box plots for odor categories are given in Figure 5 and Table S3. Correlation heat maps and their distributions are presented in Figure S5 and Figure S6. According to the analysis, cluster-14 (garlic-like, 0.479±0.128, $p<0.0001$) and cluster-6 (spicy-like, 0.443±0.0674, $p<0.001$) had higher internal correlations than the other odor categories, which indicated that these unpleasant perceptions had closer internal relationships in Word2Vec space. In contrast, some odor categories, such as cluster-10 (0.187±0.136, $p<0.0001$), cluster-8 (0.213±0.165, $p<0.01$), cluster-9 (0.217±0.139, $p<0.05$), and cluster-19 (0.227±0.174, $p<0.0001$), had lower correlations than the others. As mentioned above, cluster-10, cluster-8, and cluster-19 could not be defined as any 'well-known' smell categories. Furthermore, cluster-9 was composed of multi-odor perception categories, such as herb-like and chemical-like. However, the correlations between most odor categories were lower than 0.6, which demonstrates that the co-occurrence of terms in the text was not exactly the same as that of odor semantic descriptors.

Distribution of Odor Perception Category Labels. Before calibrating the odor category identification model, we first investigated the distribution of the samples. The sample distribution for each odor category is shown in Figure S7a, which indicates that the sample sizes for the odor descriptors were clearly distinct and imbalanced. Figure S7b illustrates the statistical distribution for the number of clusters of odorants. It demonstrates that most of the odorants belonged to more than two clusters, which can be explained by the complexity and ambiguity of odor perception.

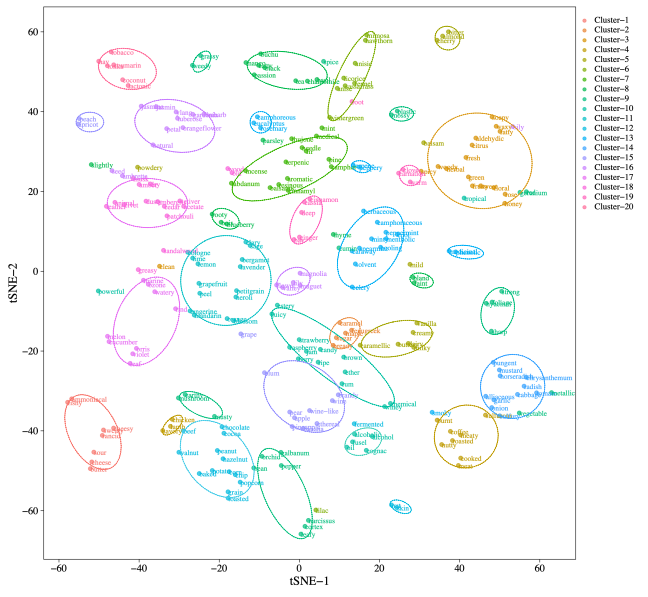

Figure 4. Odor descriptor clustering generated in the SOR space using the t-SNE method.

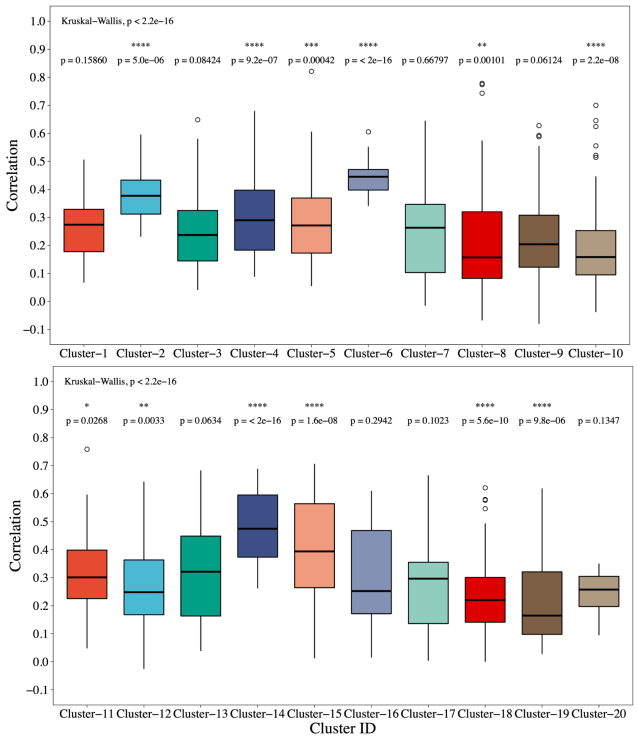

Figure 5. Correlation box plots of odor descriptor Word2Vec embeddings from odor perception categories. Results were evaluated using the nonparametric Wilcoxon signed-rank test.

Odor Perception Category Identification Models. To verify the rationality of the proposed clustering scenario, we used the molecular structure features to predict the aforementioned odor categories. To assess the potential of molecular feature extraction via four types of pre-trained CNN models, we used molecular parameters (MPs) and molecular fingerprints (FPs) to identify the clusters for odorants, applied GLVQ, and GBDT classification methods, and then compared and evaluated the results.

To calibrate the GLVQ models, the maximum number of training iterations, the number of prototypes per class were set to 5000 and 10, respectively. The overall AUC, precision, recall, and F-score of the GLVQ models under the features extracted by CNNs, molecular parameters, and molecular fingerprint data sets are shown in Figure 6, and the detailed predictions of the accuracies for each odor cluster are presented in Figure S8 and Table S4. For cluster-2, cluster-13, and cluster-15, the VGG produced better results than the other feature extraction methods. However, the features extracted by Restnet did a better job in identifying most clusters. In summary, the Restnet model produced a significantly better average AUC (0.742±0.006), precision (0.580±0.003), recall (0.691±0.005), and F-score (0.548±0.009) than the other models ($p<0.001$).

For the GBDT models, the parameters including the learning rate, max depth of trees, the fraction of features for each tree, gamma, min child weight, and subsample values were set to 0.2, 4, 0.5, 2, 0.5, and 0.5, respectively. As illustrated in Figure S9, the features extracted using GBDT with VGG showed a higher prediction accuracy for cluster-1, cluster-5, and cluster-17 than for the other clusters. In addition, Restnet did a better job of identifying cluster-7, cluster-8, cluster-10, and cluster-18. However, Densenet showed better prediction performance for most clusters. The average odor cluster identification results calibrated using the GBDT models are shown in Figure 6 and Table S4, which shows that the identification accuracy of Densenet (AUC 0.800±0.004, precision 0.595±0.004, recall 0.721±0.003, and F-score 0.570±0.007) was significantly higher than that of the other molecular feature extraction datasets ($p<0.001$).

SOR Model Comparison. When we compared the two modeling methods, we found that the GBDT had better identification accuracy than the GLVQ. In general, the GBDT run with features extracted via Desenet had the best identification performance (AUC 0.800±0.004, precision 0.595±0.004, recall 0.721±0.003, and F-score 0.570±0.007, $p<0.001$), followed by the Restnet-GBDT (AUC 0.790±0.004, precision 0.591±0.004, recall 0.719±0.004, and F-score 0.563±0.007, $p<0.001$) and Alexnet-GBDT (AUC 0.788±0.005, precision 0.589±0.004, recall 0.706±0.004, and F-score 0.562±0.007, $p<0.001$). Models trained using features extracted from molecular structure images showed higher identification accuracy (AUC 0.756±0.005, precision 0.580±0.003, recall 0.688±0.004, and F-score 0.544±0.008, $p<0.001$) than that calibrated using molecular parameters (AUC 0.522±0.003, precision 0.506±0.002, recall 0.512±0.002, and F-score 0.407±0.004, $p<0.001$). This indicates that molecular spatial structure is more highly correlated with odor perception than pure molecular parameters, which has been previously confirmed by biology experiments[30]. A similar conclusion has been noted in related work[10]. In summary, we suggest that the GBDT run with features extracted by Densenet from molecular structure images is the optimal model for identifying perception clusters of odorants.

Discussion. This paper reported an odor descriptors clustering approach aimed at testing the feasibility of defining odor categories based on the co-occurrence between odor

perceptions. We employed the BCE for describing co-occurrence relations between ODs from three odor databases. Results indicated that 265 ODs were clustered into 20 categories by HCA. Furthermore, proposed odor categories were supported by not only semantic analysis, but also the SOR. Marcelo et. al. reviewed the extant biological knowledge on olfaction to clarify the dimensionality of smell[31]. They suggested that although the human nose has over 400 olfactory receptors, the dimensionality of odor perception would be around 20 or less. Therefore, 20 odor categories proposed in the present study would be reasonable to understand the biology olfaction perception space.

For developing ML-GCO, we can train a model for odor category identification instead of OD identification, which would be easier to train. In the future, a reliable enough model would be developed as an odor perception recommendation system for the assessors of GC/O to reduce their burdens.

While our approach yielded OD clustering results based on the co-occurrence relationships, several limitations of the proposed approach should be discussed. First, abundant ODs, such as sweet and fruity, were not considered because of their ambiguities. Therefore, we need to define a metric to remove these ambiguous ODs and to identify characteristic ODs, such as coffee and vanilla. Second, the SOR model should be extended to include mixtures of odorants. Molecular structures cannot be arbitrarily blended by naive linear superposition. Thus, novel approaches and algorithms, such as molecular 3-dimensional interaction embedding, topology graph representations, and mixed MS analysis should be considered for use in extracting critical features for odor mixture presentation. We hope that, with these possible directions, our work will provide a foundation for understanding human olfaction space and finding the basis of odor perception.

■ CONCLUSIONS

In this paper, we describe a method for extracting generalized smell perceptions, termed odor categories. For clustering analysis, a metric should be defined to represent the relationship between these perceptions. In this study, we introduced BCE as a vector embedder for describing the co-occurrence relationships between ODs. The ODs of odorants were also calibrated based on the above-mentioned embedded vectors to reduce the diversity evaluation criterion in the databases. After removing infrequent ODs, cluster analyses were performed on the embedded vectors of the ODs. The results indicated that the ODs were clustered in 20 groups, and most of the ODs with similar semantic meanings were clustered in the same class.

To verify the rationality of the proposed clustering scenario, we used molecular structure features to predict the aforementioned odor categories. The high prediction accuracy (AUC is 0.8) of the SOR models demonstrated that co-occurrence-based embedded vectors may be feasible descriptors for expressing the similarity between ODs. In addition, our data demonstrate that molecular structures combined with ML methods can be adopted for odor perception cluster identification, which is a novel approach for ML-GCO.

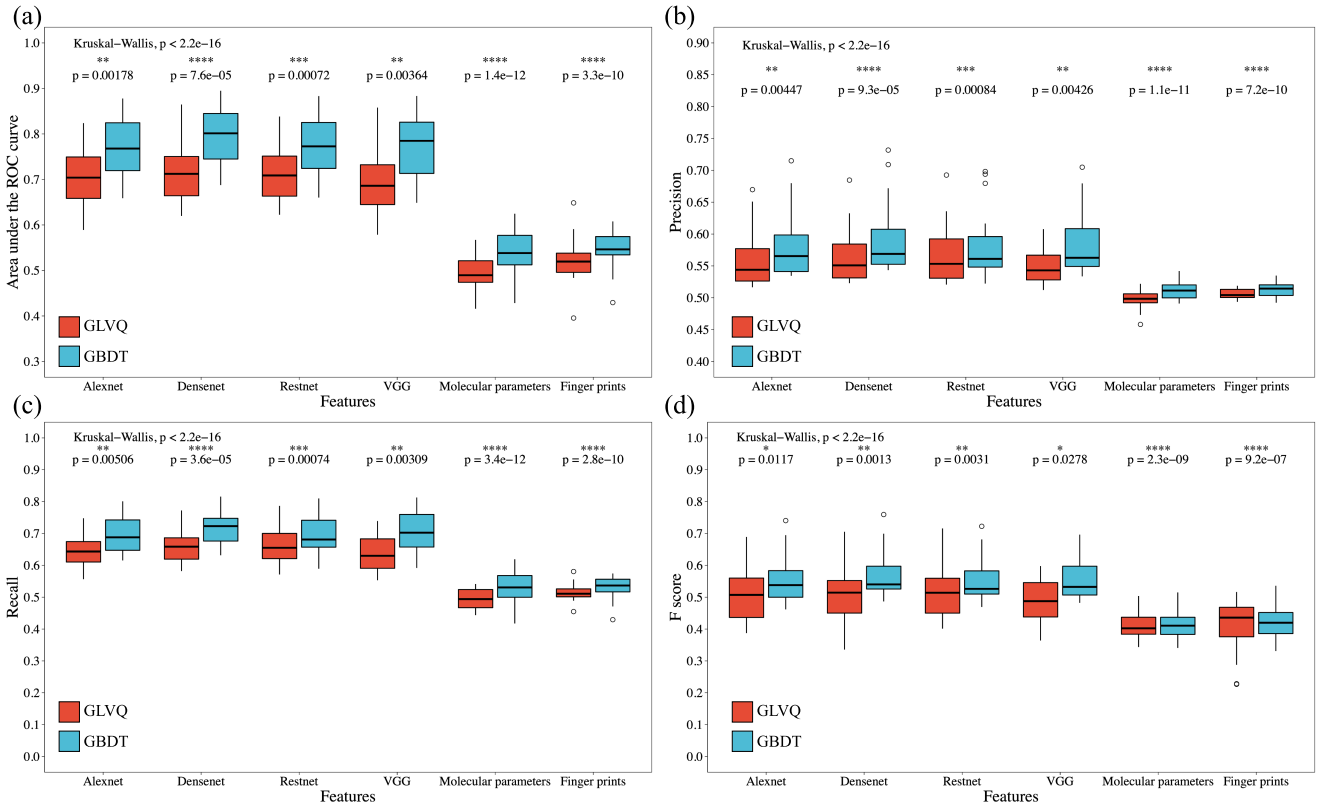

Figure 6. Comparison of average identification area under the ROC (a), precision (b), recall (c), and F-score (d) for the GLVQ and GBDT models under molecular graphic feature extractions, molecular parameters, and molecular fingerprints. Results were evaluated using the nonparametric Wilcoxon signed-rank test.

■ ASSOCIATED CONTENT

Support Information

The Support Information is available free of charge on the ACS Publications website at DOI: 00.0000/acs.anal-chem.

Detail description of algorithms and models, includes t-SNE, transductive bagging learning, model evaluation metrics, GLVQ, GDBT, supplementary figures, and tables for detailed experiment and analysis results of this work (PDF).

■ AUTHOR INFORMATION

Corresponding Authors

Liang Shang – *State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; Institute for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China;*

orcid.org/0000-0001-8369-3049;

E-mail: shangliang0225@gmail.com

Fengzhen Tang – *State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; Institute for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China;*

orcid.org/ 0000-0002-4654-9440;

E-mail: tangfengzhen@sia.cn

Authors

Chuanjun Liu – *Department of Electronics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan; Research Laboratory, U.S.E. Co., Ltd., Tokyo 150-0013, Japan*

Bin Chen – *Chongqing Key Laboratory of Non-linear Circuit and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China*

Lianqing Liu – *State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; Institute for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China*

Kenshi Hayashi – *Department of Electronics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by a grant of China Postdoctoral Science Foundation (No. 2021M703399), National Key Research and Development Program of China (No. 2020YFB13400), National Nature Science Foundation of China (No. 61803369 and 61801400), and JSPS KAKENHI Grant (No. 18H03782).

■ REFERENCES

(1) Acree, T. E. *Anal. Chem.* **1997**, *69* (5), 170A-175A. Hall, G.; Alenljung, S.; Forsgren-Brusk, U. *J Wound Ostomy Continence Nurs.* **2017**, *44* (3), 269-276. PubMed.

(2) Song, H.; Liu, J. *Food Res. Int.* **2018**, *114*, 187-198. Caporaso, N.; Whitworth, M. B.; Fisk, I. D. *Food Chem.* **2022**, *371*. Ni, R. J.; Yan, H. Y.; Tian, H. L.; Zhan, P.; Zhang, Y. Y. *Food Chem.* **2022**, *377*. Wang, Z.; Wang, Y.; Zhu, T. T.; Wang, J.; Huang, M. Q.; Wei, J. W.; Ye, H.; Wu, J. H.; Zhang, J. L.; Meng, N. *Food Chem.* **2022**, *376*. Yin, W. T.; Shi, R.; Li, S. J.; Ma, X. T.; Wang, X. D.; Wang, A. N. *Journal of Food Science* **2022**, *87* (2), 699-713.

(3) Raman, B.; Hertz, J. L.; Benkstein, K. D.; Semancik, S. *Anal. Chem.* **2008**, *80* (22), 8364-8371.

(4) Kang, J.-H.; Song, J.; Yoo, S. S.; Lee, B.-J.; Ji, H. W. *Atmosphere* **2020**, *11* (8). Liu, C.; Shang, L.; Hayashi, K. In *IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, 2019; pp 1-4.

(5) Debnath, T.; Prasetyawan, D.; Nakamoto, T. *J. Electrochem Soc.* **2021**, *168* (11), 117505. Xu, X.; Zhu, Z.; Wang, Y.; Wang, R.; Kong, W.; Zhang, J. *Commun. Nonlinear Sci. Numer. Simul.* **2022**, *109*, 106274.

(6) Gutiérrez, E. D.; Dhurandhar, A.; Keller, A.; Meyer, P.; Cecchi, G. A. *Nat. Commun.* **2018**, *9* (1), 4979. Rugard, M.; Jaylet, T.; Taboureau, O.; Tromelin, A.; Audouze, K. *PLOS ONE* **2021**, *16* (5), e0252486. DOI: 10.1371/journal.pone.0252486. Albastaki, Y. In *Artificial Intelligence Systems and the Internet of Things in the Digital Era*, Cham, 2021//, 2021; Musleh Al-Sartawi, A. M. A., Razzaque, A., Kamal, M. M., Eds.; Springer International Publishing: pp 46-56.

(7) Shang, L.; Liu, C. J.; Tomiura, Y.; Hayashi, K. *Anal. Chem.* **2017**, *89* (22), 11999-12005.

(8) Barwich, A.-S. *Cell* **2020**, *181* (4), 749-753.

(9) Nara, K.; Saraiva, L.; Ye, X.; Buck, L. *J Neurosci* **2011**, *31* (25), 9179-9191.

(10) Sharma, A.; Kumar, R.; Ranjta, S.; Varadwaj, P. K. *J. Chem. Inf. Model.* **2021**, *61* (2), 676-688.

(11) Kumar, R.; Kaur, R.; Auffarth, B.; Bhondekar, A. P. *PLOS ONE* **2015**, *10* (10), 1-19.

(12) Korichi, M.; Gerbaud, V.; Floquet, P.; Meniai, A. H.; Nacef, S.; Joulia, X. Quantitative structure-Odor relationship: Using of multidimensional data analysis and neural network approaches. In *Computer Aided Chemical Engineering*, Marquardt, W., Pantelides, C. Eds.; Vol. 21; Elsevier, 2006; pp 895-900. Pellegrino, R.; Crandall, P. G.; Seo, H.-S. *Sci. Rep.* **2016**, *6*, 18890-18890. PubMed. Wienisch, M.; Murthy, V. N. *Sci. Rep.* **2016**, *6* (1), 29308. Sharma, A.; Saha, B. K.; Kumar, R.; Varadwaj, P. K. *Nucleic Acids Research* **2021**, *50* (D1), D678-D686.

(13) Villière, A.; Guillet, F.; Prost, C. Olfactometric process: new insights in automated acquisition and data treatment. In 32nd EFFoST International Conference - Developing innovative food structures and functionalities through process and reformulation to satisfy consumer needs and expectations, Nantes, France; 2018.

(14) Sigma-Aldrich. *Merck KGaA: Darmstadt* **2016**.

(15) Arn, H.; Acree, T. E.; Contis, E. T.; Ho, C.-T.; Mussinan, C. J.; Parliment, T. H.; Shahidi, F.; Spanier, A. M. Flavornet: A database of aroma compounds based on odor potency in natural products. In *Developments in Food Science*, Vol. 40; Elsevier, 1998; p 27.

(16) Johnson, B. A.; Xu, Z.; Ali, S. S.; Leon, M. *J. Comp. Neurol.* **2009**, *514* (6), 658-673.

(17) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. *Nucleic Acids Res.* **2021**, *49* (D1), D1388-D1395.

(18) Landrum, G. *GitHub and SourceForge* **2012**, *3* (4).

(19) Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, USA, 2009; AUAI Press: pp 452-461.

(20) Li, Y.; Wang, R.; Nan, G.; Li, D.; Li, M. *Decis. Support Syst.* **2021**, *146*, 113546. Wang, C. S.; Chen, B. S.; Chiang, J. H. *Neurocomputing* **2021**, *441*, 202-213. Zhang, Q.; Ren, F. *Neurocomputing* **2021**, *440*, 365-374.

(21) He, K.; Zhang, X.; Ren, S.; Sun, J. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; pp 770-778. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017; pp 2261-2269. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. *Commun. ACM* **2017**, *60* (6), 84-90. Simonyan, K.; Zisserman, A. In *International Conference on Learning Representations (ICLR)*, 2015.

(22) Mordelet, F.; Vert, J. P. *Pattern Recogn. Lett.* **2014**, *37*, 201-209.

(23) Chen, T.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016; Association for Computing Machinery: pp 785-794. Tang, F.; Fan, M.; Tiňo, P. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32* (1), 281-292.

(24) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. In *International Conference on Learning Representations (ICLR)*, 2013.

(25) Dravnieks, A. *Science* **1982**, *218* (4574), 799-801.

(26) Schredl, M. *Int. J. Dream Res.* **2019**, *12* (1), 134-137. DOI: 10.11588/ijodr.2019.1.57845 (acccessed 2022/03/01).

(27) Snitz, K.; Perl, O.; Honigstein, D.; Secundo, L.; Ravia, A.; Yablonka, A.; Endevelt-Shapira, Y.; Sobel, N. *Chem. Senses* **2019**, *44* (4), 267-278.

(28) Gross, M. *Curr. Biol.* **2019**, *29* (14), R663-R665. Jraissati, Y.; Deroy, O. *Cogn. Sci.* **2021**, *45* (1), e12930.

(29) Snchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. *ArXiv* **2019**, *abs/1910.10685*.

(30) Pashkovski, S. L.; Iurilli, G.; Brann, D.; Chicharro, D.; Drummey, K.; Franks, K. M.; Panzeri, S.; Datta, S. R. *Nature* **2020**, *583* (7815), 253-258.

(31) Meister, M. *Elife* **2015**, *4*, e07865-e07865. PubMed.

For Table of Contents Only