Article

Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models

Xin Li ^{1,*}, Dehan Luo ¹, Yu Cheng ¹, Kin-Yeung Wong ² and Kevin Hung ²
¹ The School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China; dehanluo@gdut.edu.cn (D.L.); chengyu@gdut.edu.cn (Y.C.)

² The School of Science and Technology, The Open University of Hong Kong, Hong Kong; akywong@ouhk.edu.hk (K.-Y.W.); khung@ouhk.edu.hk (K.H.)

* Correspondence: xin_li_gdut@163.com

Abstract: Semantic odor perception descriptors, such as “sweet”, are widely used for product quality assessment in food, beverage, and fragrance industries to profile the odor perceptions. The current literature focuses on developing as many as possible odor perception descriptors. A large number of odor descriptors poses challenges for odor sensory assessment. In this paper, we propose the task of narrowing down the number of odor perception descriptors. To this end, we contrive a novel selection mechanism based on machine learning to identify the primary odor perceptual descriptors (POPDs). The perceptual ratings of non-primary odor perception descriptors (NPOPDs) could be predicted precisely from those of the POPDs. Therefore, the NPOPDs are redundant and could be disregarded from the odor vocabulary. The experimental results indicate that dozens of odor perceptual descriptors are redundant. It is also observed that the sparsity of the data has a negative correlation coefficient with the model performance, while the Pearson correlation between odor perceptions plays an active role. Reducing the odor vocabulary size could simplify the odor sensory assessment and is auxiliary to understand human odor perceptual space.

Keywords: odor perception; primary odor perception descriptor; clustering; linear regression

Citation: Li, X.; Luo, D.; Cheng, Y.; Wong, K.-Y.; Hung, K. Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models. *Appl. Sci.* **2021**, *11*, 3320. <https://doi.org/10.3390/app11083320>

Academic Editor: Albert Smalcerz

Received: 12 March 2021

Accepted: 29 March 2021

Published: 7 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The process of odor perception is more complicated than visual and auditory perception [1]. It is the result of the aggregate activations of 300 to 400 different types of olfactory receptors (ORs) [2], expressed in millions of olfactory sensory neurons (OSNs) [3,4]. These OSNs send signals to the olfactory bulb, then further to structures in the brain [5,6]. The olfactory signals are ultimately transformed into verbal descriptors, such as “musky” and “sweet”.

Compared with visual and auditory perception, odor perception is quite difficult to describe. The odor impression of human beings is affected by culture background [7], gender [8], and aging [9]. There are variabilities between people in the evaluation of odor perception. Besides, the relationships between odorants and odor perception remain elusive. Although some work has revealed some physicochemical features of the odorants, such as the carbon chain length [10] and molecular size [11], are related to human odor perception, the odorant molecules with similar structures could smell quite different. In contrast, the molecules with distinct structures could be described in an identical way [12]. In summary, odor perception is an exciting but challenging issue in the field of odor research.

Currently, semantic methods are the most widely used approaches to qualify odor perceptions in practice [13–15]. Each odor perception descriptor corresponds to one odor perception, so they are considered equal in this paper.

Semantic odor perception descriptors have played an important role in food [16,17], beverage [18,19], and fragrance engineering [20] for product quality assessment and other

commercial environments [21]. Therefore, numerous domain-specific sets of verbal descriptors for qualifying odor perception have been derived [19,22–24]. Croijmans et al. revealed a list of 146 wine-specific terms used in wine reviews [19]. The American Society for Testing and Materials (ASTM) Sensory Evaluation Committee reviewed literature and industrial sources and collected over 830 odor descriptors in use. Different from the traditional manual design approaches, Thomas et al. developed a standardized odor vocabulary in an automatic way [25]. They proposed a data-driven approach that could identify odor descriptors in English and used a distributed semantic word embedding model to derive its semantic organization. In summary, the odor vocabulary is a colossal family, which increases the difficulty of odor sensory assessment in the industries. The previous odor perception works have focused on deriving more and more verbal descriptors relating to odor perception, leading to redundancy in these odor perception descriptors. The identification of the redundant odor perception descriptors is a “blank” area. In the paper, we aim to fill this gap.

Using these odor perception vocabularies, a lot of odorant-psychophysical datasets were developed [26–28]. The well-known Dravnieks dataset [29] contains 144 mono-molecular odorants, and their expert-labeled perceptual ratings of 146 odor descriptors range from 0 to 5. The DREAM dataset [30] involves 476 molecules and their perceptual ratings of 21 odor perception descriptors labeled by 49 untrained panelists, ranging from 0 to 100. These psychological odor datasets contain exuberant information on the relations between different odor perceptions. Based on these odorant-psychophysical datasets, Substantial efforts have focused on the odor perception space and the relations among different odor perceptions.

There are two main viewpoints on the dimensions of the odor perception space: low dimension and high dimension. For the low dimension viewpoint, Khan et al. found that pleasantness was the primary dimension of the odor perceptual space by principal component analysis (PCA) [31], and this finding was validated by a series of studies [32–34]. However, the degree of pleasantness is too coarse to describe the details of odor perception. Jellinek developed an olfactory representation referred to as odor effects diagram with two fundamental polarities: erogenous versus antierogenous (refreshing) and narcotic versus stimulating [35]. Zarzo and Stanton obtained similar results to the odor effects diagram by applying PCA to two public databases [20]. For the high dimension viewpoint, Mamlouk and Martinetz reported that an upper bound of 68 dimensions and a lower bound of at least 32 Euclidean dimensions for the odor space was revealed by applying multidimensional scaling (MDS) on a 171×171 dissimilarity matrix derived from Chee-Ruiter data originated from Aldrich Chemical Company’s catalog [36]. Tran et al. implemented the Isomap algorithm to reduce the dimensions of human odor perceptions for predicting odor perceptions [37]. Castro et al. identified ten primary dimension axes of the odor perception space by using non-negative matrix factorization (NMF), in which each axis was a meta-descriptor consisting of a linear combination of a handful of the elementary descriptors [38]. In short, there is no consistency in the structure of the odor perception space. Moreover, all of these works are based on the transformation approaches, and each dimension comprises multiple-element odor perceptions. Different from that, We propose a selection mechanism to narrow down the numbers of odor perception descriptors. Based on this selection mechanism, every odor perception descriptors’ meaning remains, whose meaning is easily understood. Hence, the odor experts only need to rate much fewer odor perception descriptors so that the odor assessment workload is alleviated dramatically.

For the relations between different odor perceptions, due to the linguistic nature of odor perception descriptors, natural language processing (NLP)-based approaches have been introduced [25,39–42]. Debnath and Nakamoto implemented a hierarchical clustering based on the cosine similarity matrix, which was calculated from the 300-dimensional word vectors of 38 odor descriptors [43]. However, it mainly focuses on the prediction of odor perceptions. The reduced outputs are more intangible and harsher than the originals, and the details of the original perception information were lost. Li et al. investigated the

internal relationship between the 21 odor perceptual descriptors of the DREAM dataset and found that pleasantness was inferred from “sweet” precisely, which implied that some of the odor perceptions overlapped with others. Nonetheless, 21 descriptors are insufficient to describe the entire odor perception [44]. In summary, the relationship between these odor perceptions has not been thoroughly investigated.

To address the issues of reducing the odor vocabulary, inspired by the conception of primary odorants [45], the chemicals which the specific anosmias are mapped, we proposed the definition of primary odor perceptual descriptors (POPDs), defining as the smallest set of odor perception descriptors. The perceptual ratings of the other odor perception descriptors, namely non-primary odor perception descriptors (NPOPDs), could be inferred or predicted precisely from those of the POPDs. It should be noted that the POPDs only relate to human odor perception and have no relation with the odorants. To identify the POPDs and NPOPDs, we proposed a novel selection mechanism based on machine learning, as described in Figure 1.

Figure 1. The identification framework of primary odor perceptual descriptors (POPDs). It comprises the disjunctive model and the validation model. The disjunctive model is implemented to separate the Clustering-Center Odor Perception Descriptors (CCOPDs) and Non-Clustering-Center Odor Perception Descriptor (NCCOPDs). The NCCOPDs are the undetermined descriptors. The validation model is applied to divide the undetermined descriptors into predictable and unpredictable odor perception descriptors. The CCOPDs and unpredictable odor perception descriptors are identified as the POPDs, and the predictable odor perception descriptors as the NPOPDs.

The identification framework of POPDs comprises the disjunctive and validation models. The disjunctive model is employed to identify the Clustering-Center Odor Perception Descriptors (CCOPDs) and Non-Clustering-Center Odor Perception Descriptor (NCCOPDs). The CCOPDs are part of the POPDs. The NCCOPDs are also called undetermined odor perception descriptors, which may be POPDs or NPOPDs. The validation model is implemented to distinguish the POPDs and NPOPDs in NCCOPDs, whose inputs are the CCOPDs and the outputs are the NCCOPDs. If the perceptual ratings of the NCCOPDs could be predicted precisely by those of the CCOPDs, they are predictable and regarded as redundant. Therefore, they are identified as the NPOPDs. The NCCOPDs, which could not be predicted precisely, are unpredictable and identified as the POPDs together with the CCOPDs.

The identification framework proposed in this paper is essentially a selection strategy. No transformation is made on these odor perception descriptors, so the POPDs are a subset of all the odor perception descriptors, and their original meanings retain.

To summarize, all of the odor perception descriptors fall into two categories: POPDs and NPOPDs. Moreover, the POPDs include CCOPDs and unpredictable odor perception

descriptors, as shown in Figure 2. It should be noted that the NPOPDs are inferred from the CCOPDs only. Besides, the identification framework also reveals the complicated relations between these odor perceptions. The POPDs can be used as the axes to construct a quasi-odor perception space to describe human odor perception. Therefore, it will shed light on the research of odor perception space.

Figure 2. The classification of odor perception descriptors (OPDs). All OPDs fall into two categories: POPDs and NPOPDs, and the POPDs include CCOPDs and unpredictable odor descriptors. The perceptual ratings of NPOPDs could be inferred from the CCOPDs only.

Specifically, the main contributions of this work are as follows.

- (1) Unlike the current work focusing on developing more odor perception descriptors, we propose the task of narrowing down the number of the odor vocabulary and a novel definition of POPDs.
- (2) Different from the transformation approaches, a selection mechanism based on machine learning is proposed to identify the POPDs and NPOPDs. It demonstrates that dozens of odor descriptors are redundant, which could be removed from the odor vocabulary. The relationship between POPDs and NPOPDs is revealed by the disjunctive model, which provides the mapping functions between the odor perception descriptors.
- (3) The effects of the sparsity of odor perception data and the correlations between odor perceptions on the predicting model's performance are investigated, respectively.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods. The experiment and results are discussed in Section 3. The discussion is given in Section 4. Finally, Section 5 concludes this paper.

2. Materials and Methods

2.1. Data and Description

Compared with the other odor perceptual datasets, the structure of the Dravnieks psychophysical dataset is much denser [46], which is beneficial for the identification of POPDs. Therefore, the Dravnieks dataset is adopted in this work. It contains 144 mono-molecule odorants and their corresponding perceptual ratings of 146 odor perception descriptors, forming a 144×146 matrix [29]. Each column of this matrix is a 144-dimensional odor perception descriptor vector, representing the odor perception descriptors. Each row of this matrix is a 146-dimensional molecular vector, representing the odorants or molecules. These odor descriptors were selected by ASTM Sensory Evaluation Committee, which reviewed the literature and industrial standards for odor evaluation. A large evaluation

team composed of 150 experts evaluated the perceptual ratings ranging from 0 to 5. Therefore, compared with other odor psychophysical datasets, these data have a high degree of consistency and stability. The Percent applicability (PA) scores of these odorants were used for POPDs identification [13].

Considering that sparsity is a significant characteristic of odor psychophysical data, the sparsities of odor perception descriptors and molecules are calculated. The sparsity of each molecule is defined as:

$$SD_m = \frac{n_{mzero}}{n_m}, \quad (1)$$

where n_{mzero} is the count of perceptual ratings with a value of 0 in each 146-dimensional molecular vector. n_m is the total number of odor perception descriptors, with a value of 146.

The sparsity of each odor perception descriptor is given as follows:

$$SD_d = \frac{n_{dzero}}{n_d} \quad (2)$$

where n_{dzero} is the count of elements with a value of 0 in each 144-dimensional odor perception descriptor vector, and n_d is the total number of odor molecules, with a value of 144.

The sparsity for the Dravnieks dataset is given as:

$$SD_{data} = \frac{N_d}{N} \quad (3)$$

where N_d is the count of perceptual ratings taking the value of 0 in the 144×146 perceptual matrix, and N is the total number of elements of the dataset, i.e., 144×146 .

The 146×146 correlation matrix of odor perception descriptors is calculated. The Pearson correlation coefficient between every two 144-dimensional odor perception descriptor vectors is formulated as follows:

$$\rho_{ij} = \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{(\|x_i - \bar{x}_i\|)(\|x_j - \bar{x}_j\|)}, \quad (4)$$

where x_i, x_j are the 144-dimensional odor perception descriptor vectors, and $\bar{x}_i, \bar{x}_j$ are the means of x_i, x_j , respectively.

2.2. POPDs and the Identification Frameworks

As shown in Figure 1, there are two parts in the identification framework: disjunctive model and validation model. The disjunctive model is employed to obtain the optimal inputs of the validation model, which are used to predict as many as possible NCCOPDs precisely. The validation model is conducted for the prediction of NCCOPDs, distinguishing the predictable and the unpredictable ones. The mapping functions between the CCOPDs and the NPOPDs are established.

Inspired by [38,41], the odor perceptions cluster in the odor perception space, and the clustering centers are the representatives of the clusters. Therefore, the clustering algorithms with an updating strategy are adopted as the disjunctive models. Considering that ball-shaped clusters' clustering centers could represent the other points in the same cluster better than those of irregular-shaped clusters, the K-means algorithm and Gaussian mixture model (GMM) with an updating strategy for identifying CCOPDs are used as the disjunctive model. These two kinds of disjunctive models are called the K-means-disjunctive model and GMM-disjunctive model, respectively.

It should be noted that the CCOPDs obtained from the disjunctive models with an updating strategy are not the clustering centers determined only by the clustering algorithms. Therefore, the conclusion drawn from any clustering algorithm is no longer suitable for these CCOPDs. In other words, the disjunctive models' results could not be explained by the clustering theory, and these results could not reveal the structure of

odor perception space and any relationship between these odor perception descriptors. The disjunctive model is only used to identify some critical odor descriptors with a high probability.

To validate which NCCOPDs could be inferred or predicted precisely by the CCOPDs, three classical and stable linear regression models, namely Ridge regression, LASSO regression, and ElasticNet regression, are implemented as the validation model.

Consequently, there are six combinations for the identification frameworks: K-means-Ridge, K-means-LASSO, K-means-ElasticNet, GMM-Ridge, GMM-LASSO, and GMM-ElasticNet.

2.2.1. Disjunctive Models

The mechanism of identifying the cluster-centering odor perception descriptors is illustrated in Figure 3. Specifically, these two clustering algorithms, K-means and GMM, are conducted to obtain clustering centers. Then, CCOPDs are derived from the clustering centers by an updating strategy based on a correlation threshold (CT) to prevent highly correlated odor perception descriptors from being selected as the CCOPDs.

Figure 3. The identification of the cluster-centering odor perception descriptors by the disjunctive model. The clustering algorithm is conducted to obtain clustering centers. Then, CCOPDs are derived from the clustering centers by an updating strategy under the constraints of frequency and correlation threshold (CT).

K-Means and GMM Clustering Algorithms

For K-means, Euclidean distance is used as a similarity metric. The cost function is:

$$J(\mu_1, \mu_2, \dots, \mu_k) = \sum_{j=1}^K \sum_{i=1}^{N_j} \|x_i - \mu_j\|^2 \quad (5)$$

where K is the number of the clustering centers, N_j is the number of samples belonging to the j_{th} cluster, x_i is the 144-dimensional odor perception descriptor vector, and $\mu_1, \mu_2, \dots, \mu_K$ are the clustering centers, respectively.

By calculating the partial derivative of the cost function to minimize it, the clustering centers are formularized as follows:

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N_j} x_i \quad (6)$$

where N_j and x_i are defined as above, and N is the total number of odor perception descriptor vectors.

Another disjunctive model is based on GMM. The GMM distribution is written in the form of Gaussian linear superposition:

$$p(x) = \sum_{i=1}^K \pi_i \mathcal{N}(x | \mu_i, \Sigma_i), \quad (7)$$

where $\mathcal{N}(\cdot)$ represents the Gaussian distribution. x is the 144-dimensional odor perception descriptor vector, K is the number of the Gaussian distributions. π_i is the probability of the

i_{th} Gaussian distribution, and μ_i, Σ_i are the mean and covariance matrix of the i_{th} Gaussian distribution, respectively.

The negative log-likelihood functions of GMM is given by:

$$L_{\pi, \mu, \Sigma}(x) = \sum_{j=1}^N \log \left(\sum_{i=1}^K \mathcal{N}(x_j | \mu_i, \Sigma_i) \right), \quad (8)$$

where $x, K, \pi_i, \mu_i, \Sigma_i$ are defined as above, and N is the total number of samples, and the value is 144.

The parameters π_i, μ_i, Σ_i are determined by the Expectation Maximization (EM) algorithm and are given as:

$$\begin{cases} \mu_i = \frac{1}{\sum_{j=1}^N \gamma(j, i)} \sum_{j=1}^N \gamma(j, i) x_j \\ \Sigma_i = \frac{1}{\sum_{j=1}^N \gamma(j, i)} \sum_{j=1}^N \gamma(j, i) (x_j - \mu_i)(x_j - \mu_i)^T \\ \pi_i = \frac{\sum_{j=1}^N \gamma(j, i)}{N} \end{cases} \quad (9)$$

where x_j is j_{th} sample, K, N are defined as above, and $\gamma(j, x)$ is defined as follows:

$$\gamma(i, j) = \frac{\pi_i \mathcal{N}(x_j | \mu_i, \Sigma_i)}{\sum_{k=1}^K \pi_k \mathcal{N}(x_j | \mu_k, \Sigma_k)}. \quad (10)$$

All the 146 odor perception descriptors vectors are used to identify the clustering centers for both clustering algorithms. Considering the size of the Dravnieks dataset, the number of clustering centers ranges from 10 to 100. K is the number of NCCOPDs.

Updating Strategy for Obtaining CCOPDs from Clustering Centers

Due to the random initialization of these clustering algorithms, each iteration's clustering centers are not always the same. Therefore, each clustering algorithm has been implemented many times to improve the stability of the clustering results. In each execution of the clustering algorithms, after obtaining the clustering centers by K-means or GMM algorithms, the odor perception descriptor vectors closest to these clustering centers are updated to be the candidate CCOPDs. The frequencies of these candidate CCOPDs are recorded to identify the final CCOPDs. The frequency refers to the number of times an odor perception descriptor is selected as the candidate clustering-center odor perception during the executions. It should be noted that the number of candidate CCOPDs is always larger than the specified number of cluster centers. For example, if the specified number of clustering centers K is 10, the number of candidate CCOPDs could reach up to 50 due to the clustering algorithms' randomness. The frequencies of the candidate CCOPDs are quite different. For instance, the frequency of "fragrant" selected as the candidate CCOPDs is 894, while the frequency of "pear" is only 1.

Considering that some odor perception vectors are highly correlated, they could be selected as the candidate CCOPDs by a similar probability. Hence, it is not reasonable to use only the frequencies as the metric for determining the CCOPDs from the candidate ones. It would make the selected CCOPDs concentrate within some local vicinity in odor perceptual space. The complementary information among different clusters could not be

captured, and the information is essential for identifying the POPDs. To this end, a CT is set as a constraint.

To be specific, the candidate clustering-center odor perception descriptor with the highest frequency is selected as the first clustering-center odor perception descriptor. The other clustering-center ones are chosen as those with high frequencies and satisfy the CT constraint; that is, the correlation between them and all the selected CCOPDs are below the CT. In a word, the top K candidate clustering-odor perception descriptors with the highest frequency, whose correlation coefficients with other candidate ones are lower than the CT, are identified as the final CCOPDs.

For these two disjunctive models based on clustering algorithms, the Pearson CT or the upper bound is searched in set {0.5, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.0}. Although the number of clustering centers ranges from 10 to 100. It should be emphasized that when the CT is low, the number of clustering-centers odor perception descriptors could not reach up to 100. For example, the maximum number of CCOPDs obtained from GMM with the CT of 0.6 is 57. It is because the candidate CCOPDs could not meet the constraint of the CT.

The procedures of the disjunctive model are summarized as follows:

- (1) the clustering algorithms are implemented to determine the clustering centers;
- (2) then the odor perception descriptors closest to these clustering centers are chosen as the candidate CCOPDs;
- (3) Finally, both the frequency and CT are set as the constraints to select the CCOPDs from the candidate ones.

The updating strategy for identifying the CCOPDs is illustrated in Figure 4.

Figure 4. The updating strategy for identifying the CCOPDs from the clustering centers. After obtaining the clustering centers by the clustering algorithm, the odor perception descriptors closest to these clustering centers are chosen as the candidate CCOPDs. Then, under the constraints of frequency and CTs, the CCOPDs are identified from the candidate CCOPDs.

Besides, to qualify the coverage ability of the set of the CCOPDs with different CTs, the coverage coefficient (CC) and the augmented coverage coefficient (ACC) are introduced.

CC is defined as follows:

$$CC = \frac{\sum_{i=1}^{m_c} n_i}{\sum_{j=1}^m n_j}, \quad (11)$$

where n_i, n_j are the frequencies for i_{th}, j_{th} cluster-center odor perception descriptors selected in these executions, respectively. The maximum value for n_i and n_j is 1000. m_c is the number of the clusters specified. m is the total number of all the candidate clustering-centers odor

perception descriptors. Therefore, the denominator is the multiplication of the specified number of clusters and the execution times.

ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{m_c} n_{ic}}{\sum_{j=1}^m n_j} \quad (12)$$

where n_{ic} is the total frequency of all the candidate CCOPDs whose correlations with the i_{th} CCOPDs are above the CT. The frequency of any candidate CCOPD can only be calculated once.

2.2.2. Validation Models

As mentioned above, validation models identify the predictable and undetermined odor perceptual descriptors from the undetermined ones obtained from the former disjunctive model. The perceptual ratings of the CCOPDs are feed into the validation model as the inputs, and the perceptual ratings of the undetermined odor perceptual descriptors are as the outputs. Therefore, this validation task is a multi-output regression problem in the machine learning community [47]. In order to solve this multi-output regression problem, an assembly model composed of multiple independent regression models can be used, where each model is applied to each output variable or a single regression model with all target variables as output is used. The former approach is adopted in this paper due to its flexibility in fitting the data. Due to this dataset's moderate size, the classical and stable linear regression models have implemented: Ridge regression, LASSO regression, and ElasticNet regression. Therefore, a separate linear regression function is employed to predict one undetermined odor perception in the assembly model.

The hypothesis for a single linear regression task is given as follows:

$$y(x, w) = w^T x + w_0 \quad (13)$$

where x is the K -dimensional vector of the ratings for the clustering-centers odor perception descriptors, $x \in \mathbb{R}^K$, K is the number of clusters, w is the weight vector, and w_0 is the bias. The hypothesis is shared among three linear regression models. The difference lies in the cost function, which causes the parameters of these models to be different. Using square error to optimize, the cost function of the Ridge regression model is expressed as:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (y(x_i, w) - y_i)^2 + \lambda \sum_{j=1}^K w_j^2 \quad (14)$$

where y_i is the real label of the output, i.e., the perceptual ratings of a NCCOPD. m is the total number of training samples, and λ is a hyperparameter to balance the model complexity and avoid overfitting.

The cost function of LASSO regression model is given as:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (y(x_i, w) - y_i)^2 + \lambda' \sum_{j=1}^K |w_j| \quad (15)$$

where λ' is the hyperparameter for LASSO model, functioning the same as the hyperparameter of Ridge model.

The cost function of ElasticNet model is expressed as:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (y(x_i, w) - y_i)^2 + \lambda'' \left(\rho \sum_{j=1}^K w_j^2 + (1 - \rho) \sum_{j=1}^K |w_j| \right) \quad (16)$$

where λ'' and ρ are the hyperparameters.

The Pearson correlation coefficients between the predicted values and the actual labels are used as the performance metric for each model based on representative literature in odor perception prediction [44,48,49]. A quarter of the samples is used for testing. the other samples are used for training and validation, and ten-fold cross validation are implemented. λ for Ridge regression model is searched in set {0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100}. λ' for LASSO regression model is searched in set {0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 200, 400}. For ElasticNet regression model, λ'' is searched in set {0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100}, and ρ is searched in set {0.1, 0.2, 0.5, 0.7, 0.9}. The training and testing of all the linear regression models are implemented 30 times, then the predicting performances of the testing set in all the 30 runs are averaged as the final predicting performance.

2.3. Remark

As mentioned above, there are six combinational models in total: GMM-ElasticNet, GMM-Ridge, GMM-LASSO, K-means-ElasticNet, K-means-Ridge, and K-means-LASSO. The CT and the number of clusters K are the hyperparameters of each combinational model and have a substantial impact on the identification of POPDs and NPOPDs.

We aim to find the most miniature set of POPDs, that is, to identify the maximum number of NPOPDs. Therefore, The Grid search method is adopted to find the optimal hyperparameters K and CT for each combinational model. The results of POPDs identification by these models are recorded. By comparing the identification results of these combinational models, the one with the largest number of NPOPDs, i.e., the smallest set of POPDs, is the final winner, and the other combinational models are the baselines.

3. Experimental Results

3.1. Statistical Analysis of Odor Perceptual Data

The statistics of the dataset could affect model performance. Therefore, two kinds of statistics of odor perceptual data were analyzed: sparsity and Pearson correlation matrix. The prominent characteristic of odor perceptual data is sparsity. If most elements of an odor perceptual vector are 0, it will affect the chance of becoming a clustering-center odor perception descriptor and the accuracy of its prediction. The sparsity for the Dravnieks dataset is 26.71%, which is much denser than the other odor perceptual datasets [46]. The average perceptual rating of all odor descriptors is 3.26. The maximum sparsity of an individual odor perceptual descriptor is 81.94%, which corresponds to “fried chicken”, and the minimum is 0, which corresponds to “aromatic”, “woody, resinous”, “sharp, pungent, acid”, “heavy” and “warm”. The detailed distributions of sparsity for odor descriptors are shown in Figure 5a. The numbers on the horizontal axis represent the 146 odor perception descriptors applied in the Dravnieks Dataset (please see Supplementary S1).

Figure 5. The statistics of the Dravnieks odor perceptual dataset. The numbers on the horizontal axis in (a) and (c) represent the 146 odor perception descriptors. The numbers on the horizontal axis in (b) represent the 144 odorant molecules.

In particular, the sparsity range of more than 85% of odor perception descriptors is 0 to 0.5. The sparsity of recapitulatory odor perceptual descriptors, such as “sweet”, “perfumery”, “floral”, “fragrant”, “paint”, “sweaty”, “light”, “sickening”, “cool, cooling”, “aromatic”, “sharp, pungent, acid”, “chemical”, “bitter”, “medical”, “warm”, and “heavy”, are below 0.1.

The distribution of sparsity of molecules is reported in Figure 5b. The numbers on the horizontal axis represent the 144 odorants of the Dravnieks Dataset (please see Supplementary S2). The sparsity of every molecule is lower than 0.5. The results shown in Figure 5a,b demonstrate that the Dravnieks dataset is much dense.

Pearson correlation coefficients between every two 144-dimensional odor perception descriptor vectors were calculated, as shown in Figure 5c. The numbers 1 to 146 on the horizontal axis represent 146 odor perception descriptors (please see Supplementary S1). It means that some odor perceptions have relatively high correlation coefficients with other odor perceptions for the correspondence between numbers and odor perception descriptors. Therefore, these odor perceptions might be redundant and could be inferred from other odor descriptors.

3.2. Stability Analysis of the Identification Framework for POPDs and NPOPDs

The stability analysis of the proposed identification framework's results is conducted. This framework comprises the disjunctive model based on clustering algorithms and the validation model based on multi-output regression models. The multi-regression model is one of the three classical regression models: Ridge, LASSO, ElasticNet model, so the results of the validation model is stable, and the stability of this framework only relies on the disjunctive model. The stability of the disjunctive model's results is affected by the random initialization of the clustering algorithms. To this end, the clustering algorithms are executed many times to obtain the candidate odor perception descriptors. The CCOPDs are obtained by an updating strategy based on CT and frequency constraints to ensure the disjunctive model results' stability, as described in Section 2.2.1.

The fewer the number of clusters is, the greater the fluctuation of the selected CCOPDs are. A higher CT could lead to the greater flexibility of CCOPDs. Hence, the stability of 10 CCOPDs with a CT of 1.0 was investigated.

Precisely, the clustering algorithms were executed for 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 times to obtain the candidate CCOPDs. The CCOPDs were then derived by the updating strategy based on the constraints of CT and frequency. For each of the ten numbers of the execution times for the clustering algorithms, the disjunctive model was implemented 30 times. The results are shown in Figure 6.

It can be observed that as the number of the execution times of the clustering algorithms increases, the results of CCOPDs converge, indicating the effectiveness and stability of the disjunctive model. When the number of execution times of clustering algorithms reaches up to 1000, the results of the 30 implementations of the disjunctive models are the same. That is, the same ten CCOPDs are selected with a probability of 1. Hence, in the rest of this paper, all the results are obtained under the 1000 executions of the clustering algorithms. It should be noticed that the clustering algorithm is only used to select out the CCOPDs. After 1000 executions of the clustering algorithm and the updating strategy, the disjunctive models' results are no longer the clustering algorithms' results.

As shown in Figure 6, 14 odor descriptors were chosen as the CCOPDs for GMM-disjunctive and K-means-disjunctive models, and 13 out of these 14 descriptors were overlapped. The singular one was “rancid” for the K-means-disjunctive model and “incense” for the GMM-disjunctive model, respectively. Nine out of 10 CCOPDs obtained in these models are the same, i.e., “fruity, citrus”, “fruity, other than citrus”, “floral”, “fragrant”, “rope”, “disinfectant carbolic”, “sharp pungent, acid”, “light”, and “warm”. The other odor descriptor is “sweaty” for GMM-disjunctive model and “cool, cooling” for K-means-disjunctive model. This subtle difference is due to the similarities between these two algorithms. K-means is a particular case of GMM with an isotropic covariance

matrix. As the number of clustering centers increases, the results of these two algorithms are still different.

Figure 6. The stability analysis of ten CCOPDs obtained from Gaussian mixture model (GMM)-disjunctive and K-means-disjunctive under different numbers of execution times of clustering algorithms. The color bar on the right indicates the probability of being selected as the clustering-odor perception descriptors. (a) The stability analysis of ten CCOPDs obtained from the GMM-disjunctive models under different numbers of execution times of the clustering algorithm. (b) The stability analysis of ten CCOPDs obtained from the K-means-disjunctive models under different numbers of execution times of the clustering algorithm.

3.3. CCOPDs Derived from Disjunctive Model

Although the number of clustering centers is set between 10 and 100, the CT will affect the actual number of CCOPDs or the clustering center number. When the CT is below 0.8, the actual number of CCOPDs cannot reach 100. The maximum numbers of the CCOPDs for each CT of these two disjunctive models are presented in Table 1.

Table 1. The maximum numbers of the CCOPDs obtained from disjunctive models for different CTs. The decimals in the top row are the CTs.

	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
K-means-disjunctive models	34	43	52	61	76	91	100	100	100	100	100
GMM-disjunctive models	36	52	57	67	79	94	100	100	100	100	100

The results of the CCOPDs of these two disjunctive models under CTs of 0.5, 0.65, 0.8, and 1.0 are presented in Figure 7 for illustration. The horizontal axis represents 146 odor perception descriptors (please see Supplementary S1), and the vertical axis represents the number of CCOPDs. The black blocks represent the CCOPDs. It is shown that some descriptors, such as “warm” and “fruity”, are always chosen as the CCOPDs. Some descriptors, such as “dill”, “woody, and resinous”, are never selected even if the cluster number reaches 100. For some other descriptors, their status of being chosen changes with the cluster numbers.

Figure 7. The CCOPDs identified by K-means-disjunctive models and GMM-disjunctive models with different CTs. (a,c,e,g) are the results of the CCOPDs identified by GMM-disjunctive models with the CTs of 0.5, 0.65, 0.8, and 1.0, respectively. (b,d,f,h) are the results of the CCOPDs identified by K-means-disjunctive models with the CTs of 0.5, 0.65, 0.8, and 1.0, respectively. The black blocks are the CCOPDs, and the yellow block the NCCOPDs. (a) The CCOPDs obtained from the GMM-disjunctive model with a CT of 0.50, (b) The CCOPDs obtained from the K-means-disjunctive model with a CT of 0.50, (c) The CCOPDs obtained from the GMM-disjunctive model with a CT of 0.65, (d) The CCOPDs obtained from the K-means-disjunctive model with a CT of 0.65, (e) The CCOPDs obtained from the GMM-disjunctive model with a CT of 0.80, (f) The CCOPDs obtained from the K-means-disjunctive model with a CT of 0.80, (g) The CCOPDs obtained from the K-means-disjunctive model with a CT of 1.0. The CCOPDs obtained from the GMM-disjunctive model with a CT of 1.0.

To investigate the coverage of these CCOPDs, CCs and ACCs for both disjunctive models with different CTs were calculated, and the results are presented in Figure 8. As shown in Figure 8a, for the GMM-disjunctive model, when the clustering center number is smaller than 30, both CC and ACC increase quickly. For the ACC, when the CTs are above 0.65, the converging rates of ACCs slow down. 0.65 is the critical CT to make the ACCs converge to 1 at high speed. When the CTs are below 0.8, CC will decrease slightly as the clustering center number increases. The level of 0.8 happens to be the critical CT for the number of cluster centers to reach 100, as shown in Table 1. Similar conclusions can be

drawn from the K-means-disjunctive model, as shown in Figure 8b. It implies the CCs could affect the results of the disjunctive model.

Figure 8. The results of the coverage coefficients (CCs) and the augmented coverage coefficients (ACCs) of GMM-disjunctive models and K-means-disjunctive models with different CTs. (a) The CCs and ACCs of the GMM-disjunctive model with different CTs. (b) The CCs and ACCs of the K-means-disjunctive model with different CTs.

3.4. Performance of Combinational Models and POPDs Identified

A predicting metric threshold must be set to discriminate the predictable and unpredictable odor perception descriptors from the NCCOPDs. If the predicting performance of the NCCOPDs is above the predicting metric threshold, it will be classified as the predictable odor perception descriptors. Otherwise, it is unpredictable and will be identified as a POPDs. The numbers of NCCOPDs, namely the predictable odor perception descriptors, determined by these combinational models under different predicting metric thresholds are shown in Figure 9.

Specifically, the grid search approach was applied in the six combinational models respectively for the hyperparameter K, i.e., the number of clustering centers. The maximum number of predictable odor perception descriptors with different CTs under different predicting metric thresholds are obtained, as shown in Figure 9. The horizontal axis represents the other hyperparameters of the disjunctive model, namely, the CTs. As shown in Figure 9a–c, 0.7, 0.8, and 0.9 are the predicting metric thresholds, respectively.

For each performance metric threshold, the champion model is the one obtaining the largest number of predictable odor perception descriptors among the six combinational models. For the performance metric threshold of 0.7, there are two optimal models, one is the K-means-ElasticNet model with a CT of 0.65, and the other is the GMM-ElasticNet model with a CT of 0.55. The maximum number of predictable odor perception descriptors is 71, as shown in Figure 9a. For the performance metric threshold of 0.8, the optimal model is the K-means-ElasticNet model with a CT of 0.65. The maximum number of predictable odor perception descriptors is 51, as shown in Figure 9b. For the performance metric threshold of 0.9, there are two champions. One is the GMM-ElasticNet model with a CT of 0.8; the other is the GMM-LASSO model with a CT of 0.7. The maximum number of predictable odor perception descriptors is 22, as shown in Figure 9c. To summarize, the lower the predicting threshold, the more POPDs are obtained.

Figure 9. The maximum numbers of the predictable odor perception descriptors of the six combinational models under different predicting metric thresholds. (a) The maximum numbers of the predictable odor perception descriptors under the predicting metric threshold of 0.7. (b) The maximum numbers of the predictable odor perception descriptors under the predicting metric threshold of 0.8. (c) The maximum numbers of the predictable odor perception descriptors under the predicting metric threshold of 0.9.

As expected, the ElasticNet-based model outperforms the other regression-based models in most configurations, and the LASSO-based model is better than the Ridge-based model due to its feature selection ability. When the CTs are above 0.8, the performance of these models will drop rapidly. It shows that the clustering centers are concentrated in certain areas, and the distances among them are short. Sparse regions are not involved, and supplementary information among different clusters could not be utilized, leading to undesirable results. When the CTs are below 0.6, too many isolated odor perceptions with a little relationship with the others are chosen as clustering centers. It also leads to poor prediction performance. Consequently, the disjunctive model's CT must take an appropriate value and not too large or too small. The results show the effectiveness of the CTs on the identification framework. Besides, the high predicting metric threshold corresponds to the high CT of the disjunctive model.

Considering that prediction correlation coefficients lower than 0.7 were reported in most state-of-the-art literature, 0.8 was set as the critical value for separating the predictable and unpredictable odor perceptions. The optimal CT, the optimal number of CCOPDs, the maximum number of predictable odor perception descriptors, the optimal number of CCOPDs, and the optimal number of unpredictable odor perception descriptors for all the six combinational models with the predicting metric threshold of 0.8 are summarized in Table 2.

Table 2. The results of the six combinational models under the predicting metric threshold of 0.8. OCT is the abbreviation for optimal CT. OPD is the abbreviation for odor perception descriptor. ON is the abbreviation for optimal number. MN is the abbreviation for maximum number.

	OCT	MN of the Predictable OPDs	ON of Clusters	ON of Unpredictable OPDs
GMM-ElasticNet	0.60	49	42	55
GMM-Ridge	0.65	42	40	64
GMM-LASSO	0.65	49	43	54
K-means-ElasticNet	0.65	51	41	54
K-means-Ridge	0.65	42	41	63
K-means-LASSO	0.60	49	41	56

As shown in Table 2, the optimal CTs for all the combinational models is around 0.65, and it indicates that the internal relations among different odor perceptions are exploited exhaustively under a medium CT. Besides, the optimal numbers of CCOPDs of all the

combinational models are around 40. By comparison, the K-means-ElasticNet model with a CT of 0.65 is the final winner, referring to as the optimal configuration.

The identification results of POPDs and NPOPDs are presented in Figure 10. There are 51 NPOPDs and 95 POPDs, and the POPDs consist of 41 CCOPDs and 54 unpredictable odor perception descriptors. The 54 unpredictable odor perception descriptors could be regarded as isolated points due to their unpredictability. Moreover, the validation models mapping CCOPDs to NPOPDs, revealing the complicated relationship among these odor perceptions.

Figure 10. The results of the identification of POPDs and NPOPDs.

3.5. Relationship between Predicting Performance and Cluster Numbers for Optimal Configuration

The combination of K-means and ElasticNet regression with a CT of 0.65 is the optimal configuration. The relationship between the predicting performance and the cluster numbers K was investigated in this configuration. Some predictable NPOPDs and unpredictable POPDs were selected for this purpose. The relationship between the predicting performance of these odor perception descriptors and the number of clusters ranging from 10 to 61 was investigated. The results are presented in Figure 11.

As shown in Figure 11a, the predicting performance of some predictable NPOPDs, such as “lemon”, “orange”, and “sweet”, is relatively stable, achieving a correlation coefficient above 0.9 for all the cluster numbers. On the contrary, when the number of clusters is small, the predicting performance of the other three NPOPDs, namely “apple (fruit)”, “cologne”, and “cherry”, are inadequate. As the number of clusters increases, the prediction performance will improve, especially for “apple (fruit)”. It should be noted that the correlation coefficient 0 is actually “nan”. Due to the simplicity of these models, the

predicting values of some odor perceptions remain unchanged. Therefore, the correlation coefficients between the predicting values and the original labels would be “nan”. Similar conclusions are obtained for the unpredictable POPDs in Figure 11d–f. It should be noted that some of them are predicted with a smooth correlation coefficient, and it implies that these odor perceptions, such as “burnt candle” in Figure 11f, are singular points in the odor perceptual space and have little relationship with most other odor descriptors. In contrast, as shown in Figure 11d, the predicting performance of “cooked vegetables” and “stale tobacco smoke” increases with the number of clusters and reaches a high value above 0.8. It shows that there is a complicated relationship between these odor perceptions and others. However, to minimize the total number of POPDs, a compromise is needed to classify odor descriptors corresponding to these odor perceptions as the NPOPDs.

3.6. Effect of Data Sparsity on Model Predicting Performance

Considering that sparsity is a significant characteristic of odor perceptual data, the impact of data sparsity on model performances was investigated. The investigation was implemented in the optimal configuration, i.e., the K-means-ElasticNet regression with a CT of 0.65. Because the number of clustering centers is between 10 and 61, there are 52 models in total under the optimal configuration. The optimal model refers to the model with 41 clustering centers under the optimal configuration. Specifically, the investigation was conducted in two scenarios. One was based on the best predicting performance of each odor descriptor obtained under the optimal configuration when the number of clustering centers varied from 10 to 61. The predicting performances of all the 146 odor perception descriptors were involved. The other scenario was based on the optimal model with the optimal cluster number of 41 under the optimal configuration, and the predicting performances of the other 105 non-clustering-center odor perceptions were involved. The results are shown in Figure 12, Figure 12a for the former scenario, and Figure 12b for the latter scenario.

As shown in Figure 12a, six odor perceptions, “fruity, citrus”, “fruity other than citrus”, “floral”, “beery”, “rope”, and “warm”, were selected as the clustering centers all the time, as marked “x” in the scatter diagram. The correlation between data sparsity and the best predicting performance of all the 146 odor perception descriptors is -0.39 . In Figure 12b, the six points with a predicting performance of 0 are odor perceptions with a constant predicting value, i.e., their predicting performance metrics cannot be calculatable. The correlation between data sparsity and predicting performance of 105 NCCOPDs in the optimal model is -0.5 . These results imply that the data sparsity hurts the predicting performance, i.e., the sparser the data is, the poorer the predicting performance is.

3.7. Influence of the Pearson Correlations among Odor Perceptions on Predicting Performance

The correlations among odor perceptions are different. Some of them are highly correlated, while others are not. Hence, the influence of the correlations among odor perceptions on the predicting performance was investigated. Pearson correlation coefficients are different for every two odor perceptions, and only the large ones will significantly affect the prediction. Hence, the maximum correlation coefficients were adopted for the measurement.

Figure 11. The relationship between the predicting performance of some predictable and unpredictable odor perception descriptors and the number of clustering centers K under the optimal configuration. (a–c) The results of the predictable odor perception descriptors or the predictable NPOPDs, while (d–f) are for the unpredictable odor perception descriptors or the unpredictable POPDs. (a) The relationship between the predicting performance of the predictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these predictable odor perception descriptors are between 0.9 and 1. (b) The relationship between the predicting performance of the predictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these predictable odor perception descriptors are between 0.85 and 0.9. (c) The relationship between the predicting performance of the predictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these predictable odor perception descriptors are between 0.8 and 0.85. (d) The relationship between the predicting performance of the unpredictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these unpredictable odor perception descriptors are between 0.7 and 0.8. (e) The relationship between the predicting performance of the unpredictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these unpredictable odor perception descriptors are between 0.6 and 0.7. (f) The relationship between the predicting performance of the unpredictable odor perception descriptors and the number of cluster centers. When the number of clustering-center is 41, the predicting performances of these unpredictable odor perception descriptors are below 0.6.

Figure 12. The scatter plot of the relationship between the predicting performance and the sparsities of the odor descriptors in the optimal configuration. CSPP is the abbreviation for the correlation between sparsities and predicting performance. (a) Scatter plot of best predicting performance under the optimal configuration and sparsity of each of the 146 odor perception descriptors. (b) Scatter plot of the predicting performances and sparsities of the 105 NCCOPDs in the optimal model.

The same scenarios as in Sections 3 and 3.6 were adopted to investigate the influence of the Pearson correlations among different odor perceptions on predicting performance. The only difference in the scatter plots was that the horizontal axis represented each descriptor's maximum Pearson correlation coefficient instead of the sparsity. The results of both scenarios are shown in Figure 13. As shown in Figure 13a,b, the correlation coefficients between Pearson correlations and predicting performance are 0.52 and 0.61, respectively. These results indicate that the Pearson correlation has a positive effect on predicting performance. It indicates that if an odor perception is closely related to others, that is, in the vicinity of the other odor perceptions, it could be accurately predicted. It reveals that the distribution of odor perceptions in the odor perceptual space is not uniform. Part of odor perceptions is clustered in some vicinities, while others were isolated points.

Figure 13. The scatter plot of the relationship between the predicting performance and the maximum Pearson correlations of the odor descriptors in the optimal configuration. CPCPP is the abbreviation for correlation between Pearson correlation and predicting performance. (a) Scatter plot of best predicting performance under the optimal configuration and the maximum Pearson correlation coefficient of each of the 146 odor perception descriptors. (b) Scatter plot of the predicting performances and the maximum Pearson correlation coefficients of the 105 NCCOPDs in the optimal model.

4. Discussion

A large number of odor perception descriptors applied in industries and academic literature is a challenging issue in the odor research community, and it hampers the odor sensory assessment in food, beverage, and fragrance industries. To tackle this issue, we propose the task of reducing the amount of odor perception descriptors, and we contrive a novel selection mechanism based on machine learning to identify the POPDs and NPOPDs. The selection mechanism aims to minimize the number of POPDs by maximizing the number of NPOPDs. The validation model in this selection mechanism builds mapping functions from the CCOPDs to the NPOPDs, revealing their relationship. Moreover, it demonstrates that the sparsity and correlation between every two odor perception descriptors dramatically impact predicting performance.

Compared with the expansion of odor vocabulary in the previous work [19,23–25], we propose a shrinkage strategy in this paper. Our findings demonstrate the feasibility of reducing the size of the odor vocabulary. Although the odor perception is complicated, it could be represented by a subset of odor domain-specific semantic attributes, namely, POPDs. Just like the three primary color descriptors “red”, “blue”, and “green” for visual perception. Moreover, the NPOPDs could be recovered from a weighted linear combination of the CCOPDs, just as any color can be formed by combining three primary colors.

The framework proposed in this work is based on the assumption that there are inherent links among different odor perceptions. The identification framework proposed in this paper reveals the complicated relations between POPDs and NPOPDs to some extent. It demonstrates that some odor perception descriptors could be inferred from others, and this finding is consistent with the work in [42]. However, how to derive the general odor perception descriptors is not mentioned in [42], while we propose a disjunctive model with an updating strategy to identifying these general odor perception descriptors that could be used to infer the NPOPDs. Therefore, our work could be taken as the prequel to work in [42]. Moreover, it seems that the odor perceptions are not uniformly distributed in the odor perception space. Some of them are clustered together, while some of them are isolated points. That is, they have little to do with other odor perceptions. These findings are consistent with those in [41,44]. However, the work in [44] focuses on the relations between the missing odor perception and the known ones, and [41] aims to predict the odor perception from the mass spectrum of mono-molecular odorants.

Different from the approaches based on transformation, such as the work in [31,36,38], a selection mechanism based on machine learning is proposed in this paper. The transformation approaches mainly concentrate on exploring odor perception space, and the axes are the mixtures of several elementary odor perception descriptors. The meanings of these axes are different from any of the elementary ones. For instance, the work in [38] reveals ten axes of the odor perception space, which are the linear combination of several elementary descriptors. There are mainly six odor perception descriptors for the first basis vector, i.e., fragrant, floral, perfumery, sweet, rose, and aromatic. Any existing word could not denote the exact meaning of the mixture of the six odor perception descriptors. These axes could not be used directly for odor assessment. If we want to obtain the perceptual ratings of these axes, The perceptual ratings of its elementary odor perception descriptors must be obtained first. Therefore, the findings of these transformation-based approaches could not simplify the odor assessment. Moreover, there is overlap among the ten axes identified in [38]. For example, “fragrant” is elementary of the first three axes, and it indicates that there is redundant information among these axes. By contrast, our work focuses on reducing the odor vocabulary, and a selection strategy is adopted. Hence, the original meanings of these odor perception descriptors could be reserved. As POPDs are different from each other, there is less redundant information among them. By the fewer numbers of POPDs, the odor assessment could be simplified. Only the ratings of POPDs are needed to evaluate in the odor assessment task, and the ratings of the other odor perception descriptors could be obtained through the validation models. From this viewpoint, our work is also different from the work of [20,35], which focuses on how to develop a standard

sensory map of perfume product. The reduced number of odor perception descriptors might also be auxiliary to the odor perception space's research.

In summary, a shrink strategy of odor vocabulary is proposed. The relations between different odor perceptions are revealed by the mapping functions obtained from the validation model in this work.

Several limitations of the present work and further improvement need to be mentioned. This work mainly demonstrates the feasibility of reducing the size of odor vocabulary. The mapping functions determined by the validation model are learned automatically by "machine" and could not be explained clearly by human beings. The relationship between these odor perceptions and the odor perception space's structure remains elusive for human beings. More powerful models could be adopted to reduce the size of POPDs set further. For example, multi-task learning deep neural networks could be used as the regression model to improve predicting performance, and a smaller set of POPDs could be obtained. Due to the smoothness of the predicting performance, not all the CCOPDs are helpful for the prediction. Some feature selection algorithms may improve the predicting performance and speed up the computation. The NLP techniques could be introduced to provide supplementary information among different odor perceptions. Besides, it is possible to enlarge the odor psychological dataset to capture more and more information among different odor perceptions. More effective processing of sparsity of odor perceptual data also requires a better solution.

5. Conclusions

In this paper, we first propose a task of narrowing down the number of odor perception descriptors. Towards tackling this task, a novel definition POPD is presented, defined as the smallest odor vocabulary to profile human odor perception. Unlike the classical transformation approaches, a novel selection mechanism based on machine learning is proposed for identifying the POPDs. The POPDs are only part of the odor vocabulary, and the reduced size of odor vocabulary could greatly alleviate the burden of odor assessment. The effects of data sparsity and correlations among these odor perceptions on predicting performances were explored. It indicates that data sparsity has a negative coefficient with the predicting performance, and the correlations among odor perceptions play a positive role in odor perception prediction.

Supplementary Materials: The following are available online at <https://www.mdpi.com/2076-3417/11/8/3320/s1>.

Author Contributions: Conceptualization, X.L., D.L. and Y.C.; methodology, X.L. and Y.C.; software, X.L.; validation, X.L., Y.C. and D.L.; formal analysis, X.L. and Y.C.; investigation, K.-Y.W. and K.H.; resources, D.L.; data curation, Y.C.; writing—original draft preparation, X.L.; writing—review and editing, Y.C., D.L., K.-Y.W. and K.H.; visualization, X.L.; supervision, D.L.; project administration, D.L.; funding acquisition, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the National Natural Science Foundation of China under Grant 61571140, and in part by Guangdong Science and Technology Department under Grant 2017A010101032.

Acknowledgments: The authors would like to thank the colleagues from collaborating universities for their support.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

POPD	Primary Odor Perception Descriptor
NPOPD	Non-Primary Odor Perception Descriptor
CCOPD	Clustering-Center Odor Perception Descriptor
NCCOPD	Non-Clustering-Center Odor Perception Descriptor
GMM	Gaussian mixture model
CT	Correlation Threshold
CC	Coverage Coefficient
ACC	Augmented Coverage Coefficient
ASTM	American Society for Testing and Materials
OR	Olfactory receptor
OSN	Olfactory sensory neuron

References

- Haddad, R.; Khan, R.; Takahashi, Y.K.; Mori, K.; Harel, D.; Sobel, N. A metric for odorant comparison. *Nat. Methods* **2008**, *5*, 425. [\[CrossRef\]](#) [\[PubMed\]](#)
- Su, C.Y.; Menuz, K.; Carlson, J.R. Olfactory perception: receptors, cells, and circuits. *Cell* **2009**, *139*, 45–59. [\[CrossRef\]](#) [\[PubMed\]](#)
- Buck, L.; Axel, R. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **1991**, *65*, 175–187. [\[CrossRef\]](#)
- Wilson, C.D.; Serrano, G.O.; Koulakov, A.A.; Rinberg, D. A primacy code for odor identity. *Nat. Commun.* **2017**, *8*, 1–10. [\[CrossRef\]](#)
- Menini, A.; Lagostena, L.; Boccaccio, A. Olfaction: from odorant molecules to the olfactory cortex. *Physiology* **2004**, *19*, 101–104. [\[CrossRef\]](#)
- Si, G.; Kanwal, J.K.; Hu, Y.; Tabone, C.J.; Baron, J.; Berck, M.; Vignoud, G.; Samuel, A.D. Structured Odorant Response Patterns across a Complete Olfactory Receptor Neuron Population. *Neuron* **2019**, *101*, 950–962. [\[CrossRef\]](#)
- Chrea, C.; Valentin, D.; Sulmont-Rossé, C.; Mai, H.L.; Nguyen, D.H.; Abdi, H. Culture and odor categorization: agreement between cultures depends upon the odors. *Food Qual. Prefer.* **2004**, *15*, 669–679. [\[CrossRef\]](#)
- Doty, R.L.; Cameron, E.L. Sex differences and reproductive hormone influences on human odor perception. *Physiol. Behav.* **2009**, *97*, 213–228. [\[CrossRef\]](#)
- Seubert, J.; Kalpouzos, G.; Larsson, M.; Hummel, T.; Bäckman, L.; Laukka, E.J. Temporolimbic cortical volume is associated with semantic odor memory performance in aging. *NeuroImage* **2020**, *211*, 116600. [\[CrossRef\]](#)
- Boesveldt, S.; Olsson, M.J.; Lundström, J.N. Carbon chain length and the stimulus problem in olfaction. *Behav. Brain Res.* **2010**, *215*, 110–113. [\[CrossRef\]](#)
- Kermen, F.; Chakirian, A.; Sezille, C.; Joussain, P.; Le Goff, G.; Ziesel, A.; Chastrette, M.; Mandairon, N.; Didier, A.; Rouby, C.; et al. Molecular complexity determines the number of olfactory notes and the pleasantness of smells. *Sci. Rep.* **2011**, *1*, 1–6. [\[CrossRef\]](#)
- Sell, C. On the unpredictability of odor. *Angew. Chem. Int. Ed.* **2006**, *45*, 6254–6261. [\[CrossRef\]](#)
- Dravnieks, A. Odor quality: Semantically generated multidimensional profiles are stable. *Science* **1982**, *218*, 799–801. [\[CrossRef\]](#)
- Winter, B. Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Lang. Cogn. Neurosci.* **2016**, *31*, 975–988. [\[CrossRef\]](#)
- Kaepler, K.; Mueller, F. Odor classification: A review of factors influencing perception-based odor arrangements. *Chem. Senses* **2013**, *38*, 189–209. [\[CrossRef\]](#)
- Alegre, Y.; Sáenz-Navajas, M.P.; Hernández-Orte, P.; Ferreira, V. Sensory, olfactometric and chemical characterization of the aroma potential of Garnacha and Tempranillo winemaking grapes. *Food Chem.* **2020**, *331*, 127207. [\[CrossRef\]](#)
- Pu, D.; Duan, W.; Huang, Y.; Zhang, Y.; Sun, B.; Ren, F.; Zhang, H.; Chen, H.; He, J.; Tang, Y. Characterization of the key odorants contributing to retronasal olfaction during bread consumption. *Food Chem.* **2020**, *318*, 126520. [\[CrossRef\]](#)
- Moore, T.; Carling, C. Wine and Words: An empirical approach. In *The Limitations of Language*; Springer: Berlin, Germany, 1988; pp. 121–125.
- Croijmans, I.; Hendrickx, I.; Lefever, E.; Majid, A.; Van Den Bosch, A. Uncovering the language of wine experts. *Nat. Lang. Eng.* **2019**, *26*, 1–20. [\[CrossRef\]](#)
- Zarzo, M.; Stanton, D.T. Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. *Atten. Percept. Psychophys.* **2009**, *71*, 225–247. [\[CrossRef\]](#)
- Vega-Gómez, F.I.; Miranda-Gonzalez, F.J.; Mayo, J.P.; López, O.G.; Pascual-Nebreda, L. The Scent of Art. Perception, Evaluation, and Behaviour in a Museum in Response to Olfactory Marketing. *Sustainability* **2020**, *12*, 1384. [\[CrossRef\]](#)
- Noble, A.C.; Arnold, R.; Masuda, B.M.; Pecore, S.; Schmidt, J.; Stern, P. Progress towards a standardized system of wine aroma terminology. *Am. J. Enol. Vitic.* **1984**, *35*, 107–109.
- Lawless, L.J.; Civille, G.V. Developing lexicons: A review. *J. Sens. Stud.* **2013**, *28*, 270–281. [\[CrossRef\]](#)

24. Majid, A.; Roberts, S.G.; Cilissen, L.; Emmorey, K.; Nicodemus, B.; O'Grady, L.; Woll, B.; LeLan, B.; De Sousa, H.; Cansler, B.L.; et al. Differential coding of perception in the world's languages. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11369–11376. [CrossRef]
25. Hörberg, T.; Larsson, M.; Olofsson, J. Mapping the semantic organization of the English odor vocabulary using natural language. *Data* **2020**. [CrossRef]
26. Arctander, S. *Perfume and Flavor Chemicals*; Steffen Arctander: Montclair, NJ, USA, 1969; Volume 1.
27. The Good Scents Company. The Good Scents Company Information System. Available online: <http://www.thegoodscentscompany.com> (accessed on 6 April 2021).
28. Sigma-Aldrich, B.; Region, P.S.A. Sigma-Aldrich. *Sigma* **2018**, *302*, H331.
29. Dravnieks, A. *Atlas of Odor Character Profiles*; ASTM: Philadelphia, PA, USA, 1985.
30. Keller, A.; Gerkin, R.C.; Guan, Y.; Dhurandhar, A.; Turu, G.; Szalai, B.; Mainland, J.D.; Ihara, Y.; Yu, C.W.; Wolfinger, R.; et al. Predicting human olfactory perception from chemical features of odor molecules. *Science* **2017**, *355*, 820–826. [CrossRef]
31. Khan, R.M.; Luk, C.H.; Flinker, A.; Aggarwal, A.; Lapid, H.; Haddad, R.; Sobel, N. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *J. Neurosci.* **2007**, *27*, 10015–10023. [CrossRef]
32. Yeshurun, Y.; Sobel, N. An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. *Annu. Rev. Psychol.* **2010**, *61*, 219–241. [CrossRef]
33. Koulakov, A.; Kolterman, B.E.; Enikolopov, A.; Rinberg, D. In search of the structure of human olfactory space. *Front. Syst. Neurosci.* **2011**, *5*, 65. [CrossRef]
34. Zarzo, M. Psychologic dimensions in the perception of everyday odors: Pleasantness and edibility. *J. Sens. Stud.* **2008**, *23*, 354–376. [CrossRef]
35. Paul, J.S. *The Psychological Basis of Perfumery*; Springer Science & Business Media: Berlin, Germany, 2012.
36. Mamlouk, A.M.; Martinetz, T. On the dimensions of the olfactory perception space. *Neurocomputing* **2004**, *58*, 1019–1025. [CrossRef]
37. Tran, N.; Kepple, D.; Shuvaev, S.; Koulakov, A. DeepNose: Using artificial neural networks to represent the space of odorants. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 24 May 2019; pp. 6305–6314.
38. Castro, J.B.; Ramanathan, A.; Chennubhotla, C.S. Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE* **2013**, *8*, e73289. [CrossRef] [PubMed]
39. Iatropoulos, G.; Herman, P.; Lansner, A.; Karlgren, J.; Larsson, M.; Olofsson, J.K. The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition* **2018**, *178*, 37–49. [CrossRef] [PubMed]
40. Liu, C.; Liang Shang, K.H. Co-occurrence-based clustering of odor descriptors for predicting structure-odor relationship. In Proceedings of the 2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), Fukuoka, Japan, 26–29 May 2019.
41. Nozaki, Y.; Nakamoto, T. Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. *PLoS ONE* **2018**, *13*, e0198475.
42. Gutiérrez, E.D.; Dhurandhar, A.; Keller, A.; Meyer, P.; Cecchi, G.A. Predicting natural language descriptions of mono-molecular odorants. *Nat. Commun.* **2018**, *9*, 4979. [CrossRef]
43. Debnath, T.; Nakamoto, T. Predicting human odor perception represented by continuous values from mass spectra of essential oils resembling chemical mixtures. *PLoS ONE* **2020**, *15*, 1–13. [CrossRef]
44. Li, X.; Luo, D.; Cheng, Y.; Wong, A.K.Y.; Hung, K. A Perception-Driven Framework for Predicting Missing Odor Perceptual Ratings and an Exploration of Odor Perceptual Space. *IEEE Access* **2020**, *8*, 29595–29607. [CrossRef]
45. Amoore, J.E. Specific anosmia and the concept of primary odors. *Chem. Senses* **1977**, *2*, 267–281. [CrossRef]
46. Kumar, R.; Kaur, R.; Auffarth, B.; Bhondekar, A.P. Understanding the odour spaces: A step towards solving olfactory stimulus-percept problem. *PLoS ONE* **2015**, *10*, e0141263. [CrossRef]
47. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
48. Wu, D.; Luo, D.; Wong, K.Y.; Hung, K. POP-CNN: Predicting Odor Pleasantness With Convolutional Neural Network. *IEEE Sens. J.* **2019**, *19*, 11337–11345. [CrossRef]
49. Li, H.; Panwar, B.; Omenn, G.S.; Guan, Y. Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features. *Gigascience* **2018**, *7*, gix127. [CrossRef]