ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Regression Study of Odorant Chemical Space, Molecular Structural Diversity, and Natural Language Description

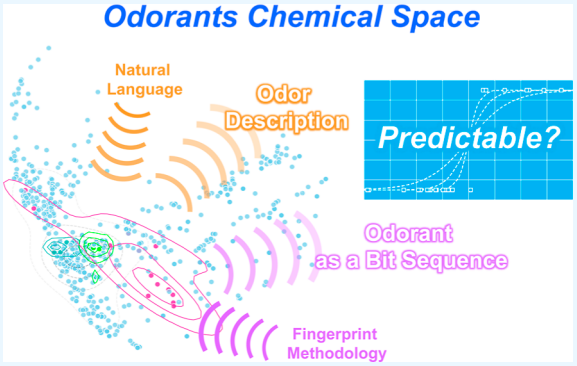
Yuki Harada,* Shuichi Maeda, Junwei Shen, Taku Misonou, Hirokazu Hori, and Shinichiro Nakamura

Cite This: *ACS Omega* 2024, 9, 25054−25062

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Odor is analyzed on the human olfactometry systems in various steps. The mapping from chemical structures to olfactory perceptions of smell is an extremely challenging task. Scientists have been unable to find a measure to distinguish the perceptual similarity between odorants. In this study, we report regression analysis and visualization based on the odorant chemical space. We discuss the relation between the odor descriptors and their structural diversity for odorants groups associated with each odor descriptor. We studied the influence of structural diversity on the odor descriptor predictability. The results suggest that the diversity of molecular structures, which is associated with the same odor descriptor, is related to the resolutional confusion with the odor descriptor.



1. INTRODUCTION

Odor is analyzed on the human olfactometry systems in various steps. The total system includes the olfactory receptor neurons in the olfactory epithelium, the glomeruli in the olfactory bulb, the peripheral olfactory system of the brain, and/or the central olfactory system of the brain. Odor shows numerical and/or non-numerical representation in these steps, respectively.[1] In this study, we propose the presence of a vector space that aids in an understanding of the whole difficult mechanism in the human olfaction.

Since odorant molecules encode the olfactory percept, several researchers have focused on mapping from molecular structures to odor perception in the current decade.[2−8] In those studies, the analysis is based on the "chemical space" of odorants. The concept of chemical space has historically been applied in a wide range of fields, such as drug discovery and functional molecular design, owing to its multiple potential applications. The chemical space is usually defined as the set of all of the possible organic compounds. Of late, the mapping from chemical structures to olfactory perception has been attracting considerable attention. However, up to today, scientists have been unable to develop a physical measure; there is not a straightforward way to link odorants and descriptors. Human olfactory intervention is an essential issue. The participation of the human sense, which is not possible to replace by a machine, is still indispensable. The relationship between the chemistry and perception of odors remains unclear; even a clue is unclear to merely digitize the extent of perceptual similarity between odorants.

For exploring the chemical space, it is self-evident that the structural diversity or variety of chemicals must be of fundamental importance. However, the "diversity" is still, to some extent, a subjective concept because of the difficulty of giving a numerical measure. There are four main strategies to approach the structural diversity that have been consistently identified in the literature: (i) appendage diversity (or building-block diversity), (ii) functional group diversity, (iii) stereochemical diversity, and (iv) skeletal (or frameworks/scaffolds) diversity.[9−11] Since the diversity is still difficult to quantify, we consider the similarity. We adopted a conventional approach of Tanimoto similarity scores as an index for structural diversity in the current study.

In this paper, we investigate the reason why mapping from chemical structures to olfactory perceptions of smell is an extremely challenging task. We propose a clue to get into an effective approach. We believe that the classification of the difficulty provides an aid to the design of new odorants. As the first step of the approach to this subject, we will present regression analysis and mapping based on the odorant chemical space. The "explanatory" odor-sensing space is a fingerprint representation of each chemical space, and the "target" odor-sensing space is odor descriptors. We will discuss the relation between the odor descriptor and each structural diversity in their chemical spaces, indicating why and how there is difficulty in characterizing and digitizing the odor.

Received: March 8, 2024
Revised: May 15, 2024
Accepted: May 24, 2024
Published: June 3, 2024

Table 1. Odor Descriptors in the Flavornet Database for the Current Study

Odor Descriptor	Num of Odorants (Frequency)	Odor Descriptor	Num of Odorants (Frequency)	Odor Descriptor	Num of Odorants (Frequency)
fruit	69	roast	19	nut	11
green	62	must	17	camphor	10
sweet	60	oil	16	coconut	10
wood	46	pungent	16	turpentine	9
herb	45	balsamic	16	cabbage	10
flower	45	rose	15	medicine	10
spice	42	fresh	15	lemon	9
fat	35	apple	14	rancid	9
sulfur	22	caramel	14	cucumber	8
citrus	22	wax	13	mushroom	8
mint	20	honey	11	soap	8
earth	20	nut	11		
alkane	20	metal	10	(other 158 descriptor)	< 8

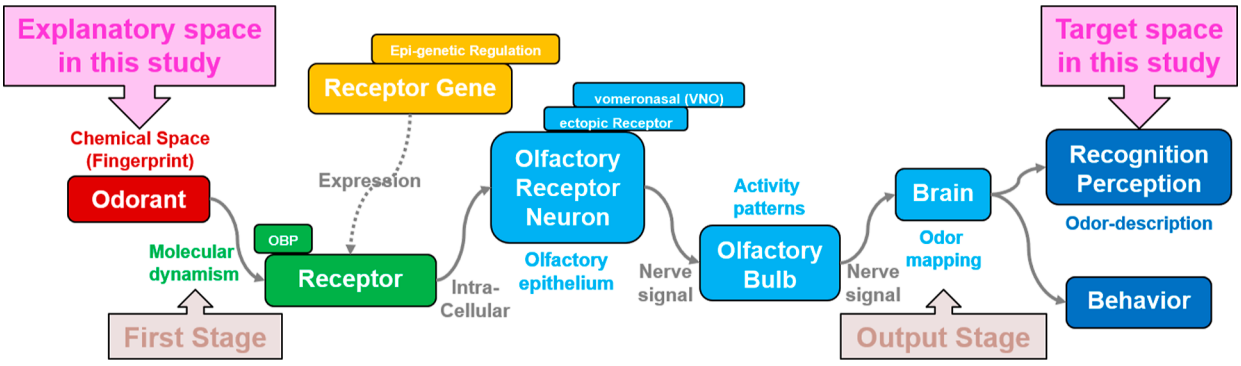


Figure 1. Odor informatization of the human olfactometry system.

2. MATERIALS AND METHODS

2.1. Odorants and Odor Descriptors in the Flavornet Database. The Flavornet database is a compilation of aroma compounds found in the human odor space.[12] At least 10,000 odorants invoke a huge number of human olfactory perceptions. In this database, odorants are arranged by chromatographic and odor descriptors. The data of the current study are collected from articles published since 1984 using GCO to detect odorants in natural products. In Table 1, we show odor descriptors and the number of odorants (molecules) extracted from the Flavornet database. The rest of the database is in the Supporting Information (S1).

Notice that the database has generally an unavoidable difficulty between odorants and odor descriptions: the main limitation in establishing the structure−odor relation is that the descriptor labeling (the word for descriptor) is vague and ambiguous in nature because the descriptors consist of natural languages. Kaeppler reported that the word for descriptor can also originate from contextual cues such as the color or verbal label.[13] When it comes to descriptors for taste, they are derived from the words indicating the taste itself. To the contrary, the descriptors for smell are derived from natural language, which requests one step of transformation, odor of sweet, odor of green, etc. The situation is especially serious when odorant molecules are collected from different sources.[6]

2.2. Fingerprints. The molecular fingerprint is a method to represent a molecule as a sequence of bits (on or off); it encodes features of the molecular structure. The molecular fingerprints are representations of chemical structures in the chemo-informatics database, invented in the early days. It is used for search and analysis, such as similarity searching, clustering, and classification. The fingerprints are used as a measure of "molecular distance" in substructure screening or as inputs for machine learning functions.

In this study, we tried four fingerprints: MACCS keys, ECFPs, Avalon fingerprints, and RD-kit fingerprints. The MACCS keys fingerprint has 166 bits structural key descriptors (a vector with 166 elements) in which each bit is associated with a SMARTS pattern.[14,15] Extended-connectivity fingerprints (ECFPs) are circular topological fingerprints designed for various wide molecular studies and structure−activity modeling.[16,17] The ECFP encodes substructure patterns from molecules on to the bit string of length 1024 (length can be varied). The Avalon fingerprint is a hashed fingerprint enumerating paths and feature classes. Similar to Daylight fingerprints, the Avalon fingerprint uses a fingerprint generator that enumerates certain paths and feature classes of the molecular graph. The RDKit fingerprint[18] is a hashed substructure/path fingerprint similar to the Daylight fingerprints.[19]

We carried out the analysis by these four fingerprints, and the essential arguments are qualitatively conserved; thus, we explain mainly the results developed by RDkit, and the results by others are also shown in Supporting Information (S1). For the similarity measures of pairs in each chemical space, Tanimoto similarity scores are used for all odor descriptors. We adopt this as an index for molecular diversity, to be discussed later.

2.3. Mapping. For comparison and exploration of the internal relations in the chemical space, it is necessary to map the complicated higher-order vector space of various fingerprints onto a low-dimensional space. We adopt the principal component analysis (PCA), which is a typical mapping widely used for exploratory data analysis and to make predictive models.[20] We plotted the first two principal components.

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

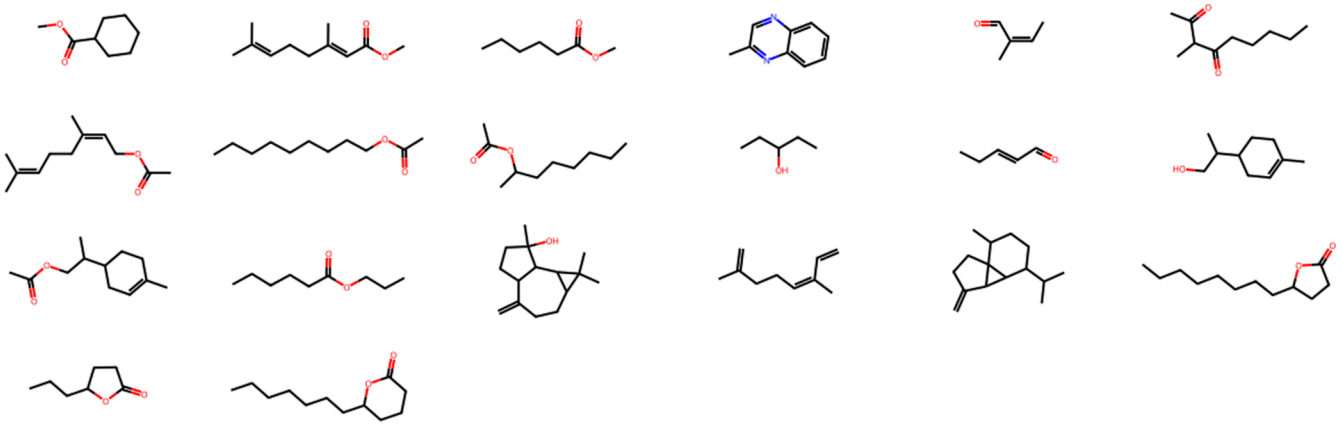


Figure 2. Molecules of the odor descriptor "fruit".

Then, kernel density estimation (KDE)[20] is applied to estimate the density of odor descriptors on the plot space. The contour lines show their obtained KDE. Dots are shown by the vertical and horizontal axes of the first two principal components. Results are shown later in Section 3.2.

2.4. Regression Model and Its Evaluation. We performed logistic regression analysis between the odorant molecules and the descriptors. The regression model was evaluated by *k*-fold cross-validation [see Supporting Information (S5)],[21] and then the average of the values computed in the loop was reported. The grid search was used for optimizing hyperparameters of this regression model using the first $k-1$ of the fold as training data. The model is trained using the first $k-1$ of folds as training data, and then the resulting model is validated on the remaining part of the data. A receiver operating characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

For example, we show the model evaluation of the odor descriptor "fruit". The 69 "fruit" chemicals were found in the Flavornet database. Among the 69 molecules, we have shown representative 20 molecules in Figure 2, and the rest of the 49 chemicals are in Supporting Information (S1). By the regression model, the resultant ROC curve of evaluation is shown in Figure 3 [area under curve (AUC) of ROC curve: 0.70, by the RDKit fingerprint]. We perform the process for all odor descriptors and for four fingerprints (Supporting Information (S1)).

As described above, the predictability was obtained by iterative *k*-fold cross-validation for the regression model for each odor descriptor.

3. RESULTS AND DISCUSSION

This study reports the regression analysis, in which the explanatory variables are fingerprint representations of odorants and the target variables are binary class labels corresponding to one odor descriptor. The objective of the regression model is to examine whether one odorant might have one odor descriptor. The descriptors are adopted from natural languages since there are not enough words to describe the smell directly. Therefore, natural languages are borrowed. It is not always evident that the borrowed language is appropriate to the smell of interest. Thus, odor descriptors in the Flavornet database are also considered to be mainly adopted from nonolfactive languages.

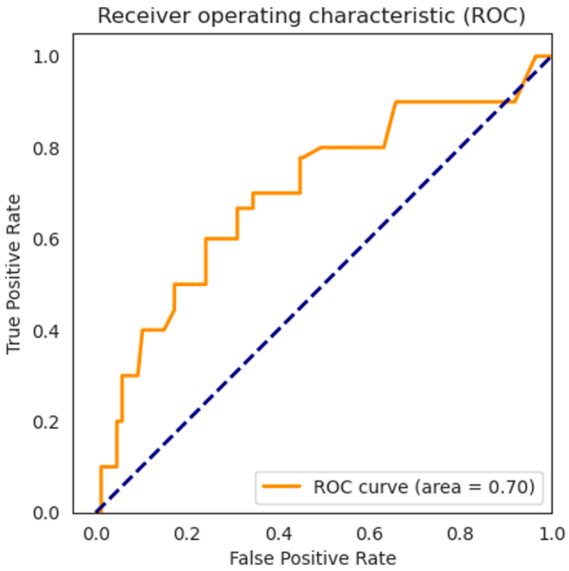


Figure 3. ROC curve for regression for the target odor descriptor: "fruit".

The frequency (number of odorant molecules) of one odor descriptor is shown in Table 1. It is noteworthy that the relation between frequency and the odor descriptors is not homogeneous. That is, some descriptors have many odorant molecules, and others have only a few molecules. There are cases in which one descriptor has only one odorant. The examples are shown in Supporting Information (S1). The descriptors with only one odorant are somehow unpopular and are the subject of future study. In the current study, we define the molecules having the same odor descriptor as the "odorants group". We investigate the relation of the descriptor to the odorants group.

Because there is not a unique definition to transfer the structural properties into an indexing number. Their similarity and diversity are only partially evaluated, and it is unclear how much they differ by number. There are some metrics to calculate the distance among molecular structures, including machine learning metrics such as the GNN-based[22] and the latest.[23,24] Tanimoto similarity scores show less discriminatory power for the chemical space than the latest metrics. However, it has been widely used and tested for a long time. We applied it for the pairwise group among the same odor descriptor as an index for structural diversity.

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

When the value is 1.0, it is the case that the similarity is very close. The average of similarity scores of each odorants group is shown in Table 2. The color in the matrix of the left side, shown with a gradation from green to white, represents the average of Tanimoto similarity scores. The lower scores are shown with darker colors (low similarity), whereas the higher scores are shown with lighter colors (high similarity). The results by other three fingerprints are in Supporting Information (S1). The arguments described above are consistent.

Table 2. AUC Values and Similarity Scores for Each Odor Descriptor Obtained by the RDKit Fingerprint; the Odorants Groups Marked as *1 and *2 Are the Examples of "High Predictability" and "Low Predictability", Respectively

Odor Descriptors		Average of Tanimoto Scores	AUC (Training)	AUC (Validation)	
fruit	69	0.200	0.897	0.705	
green	62	0.170	0.907	0.705	
sweet	60	0.121	0.911	0.619	
wood	46	0.311	0.931	0.927	
herb	45	0.184	0.933	0.754	
flower	45	0.132	0.933	0.709	
spice	42	0.183	0.943	0.828	
fat	35	0.283	0.950	0.893	
sulfur	22	0.085	0.978	0.954	
citrus	22	0.168	0.967	0.702	
mint	20	0.219	0.971	0.765	
earth	20	0.169	0.974	0.786	
alkane	20	0.859	0.997	1.000	*1
roast	19	0.144	0.971	0.896	
must	17	0.108	0.974	0.429	*2
oil	16	0.192	0.976	0.694	*2
pungent	16	0.215	0.976	0.888	
balsamic	16	0.143	0.978	0.692	
rose	15	0.321	0.985	0.983	*1
fresh	15	0.181	0.978	0.691	*2
apple	14	0.351	0.979	0.911	
caramel	14	0.170	0.983	0.724	
wax	13	0.208	0.981	0.711	
honey	11	0.247	0.983	0.726	
nut	11	0.134	0.983	0.723	
metal	10	0.226	0.985	0.828	
sweat	10	0.330	0.990	0.914	
camphor	10	0.351	0.988	0.914	
coconut	10	0.470	0.985	0.961	
cabbage	10	0.153	0.990	0.997	*1
medicine	10	0.094	0.985	0.590	
turpentine	9	0.253	0.986	0.817	
lemon	9	0.249	0.991	0.775	
rancid	9	0.223	0.986	0.882	
cucumber	8	0.409	0.993	0.958	*1
mushroom	8	0.235	0.988	0.838	
soap	8	0.346	0.988	0.901	

3.1. Cross-Validation Result of the Regression Model.

In Figure 3, we show the ROC curve of the descriptor "fruit"; for example, the other ROC curves are in the Supporting Information (S1). In Table 2, we show the heat map of the AUC variance of the ROC metric obtained by iterative *k*-fold cross-validation of the regression model for each odor descriptor. In order to avoid arbitrariness depending on a fingerprint, we carried out regression model analyses by four fingerprints. We show the results by the other three fingerprints in Supporting Information (S1). The color in the matrix of the middle position, shown with a gradation from yellow to white, represents the mean AUC in "training", while the color in the matrix of the right side, shown with a gradation from red to white, represents the mean AUC in "validation". The lower AUC values are shown with darker colors ("low predictability"), whereas the higher AUC values are shown with lighter colors ("high predictability").

We plotted the results of the AUCs of each odorants group as shown in Figure 4. The AUC value is in vertical axes, and the average of Tanimoto similarity scores, an index of structural diversity in each odorants group, is in horizontal axes. The logistic regression and the similarity scores for each odor descriptor were obtained by the sklearn package and the RDKit package, respectively. The other plots of AUCs based on ECFPs, MACCS keys, and Avalon fingerprints are shown in Supporting Information (S1).

There is a broad correlation between the AUC and structural diversity. That is, the odorants groups with a large average of Tanimoto similarity scores have narrow structural diversity. They have relatively high AUC values. It indicates that the prediction can be effective when structural diversity is narrow. It also indicates that structural diversity influences the predictability. Nevertheless, there are four or three outlier points (SU and CA, upper left of Figure 4) to be discussed later.

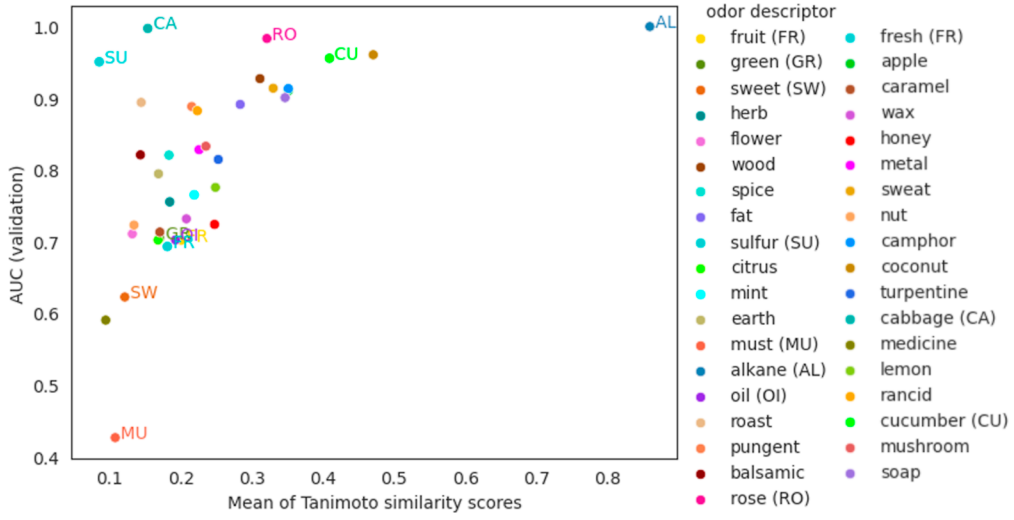


Figure 4. Correlation between AUC values and the average of Tanimoto similarity scores (RDKit fingerprint).

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

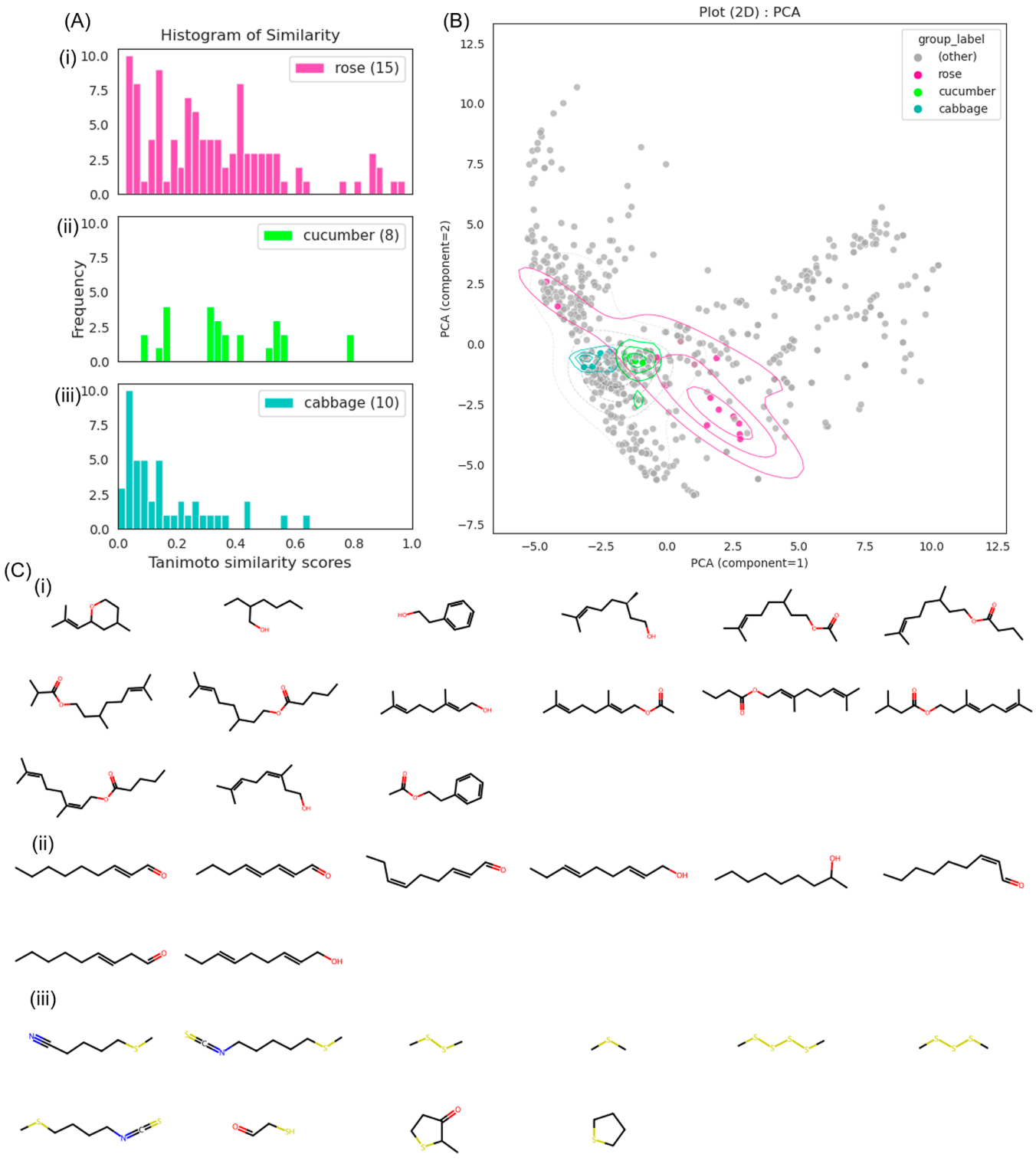


Figure 5. Structural diversity of the "high predictability odorants group"; (A) histogram of pairwise Tanimoto similarity scores in each odorants group; (B) mapping of principal component analyses based on RDKit fingerprints; and (C) molecules in the "high predictability odorants group" (i) "rose", (ii) "cucumber", and (iii) "cabbage".

In an attempt to clarify the relation thus obtained between structural diversity and the odor descriptor associated with the odorants group, we obtained the histograms of the Tanimoto similarity score. In fact, we calculated the pairwise values across all odorants groups. They are shown in Figures 5A, 6A, and 7A. They are the results of Tanimoto similarity scores by the RDKit fingerprint; the results by other fingerprints are in Supporting Information (S1). The similarity values (Tanimoto similarity scores) are in the horizontal axes and the frequency (histogram of how many pairs are included) is in the vertical axes. We will discuss the results below.

3.2. Correlation between Predictability and Structural Diversity. The descriptors "rose", "cucumber", and "cabbage" are the examples of "high predictability". These

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

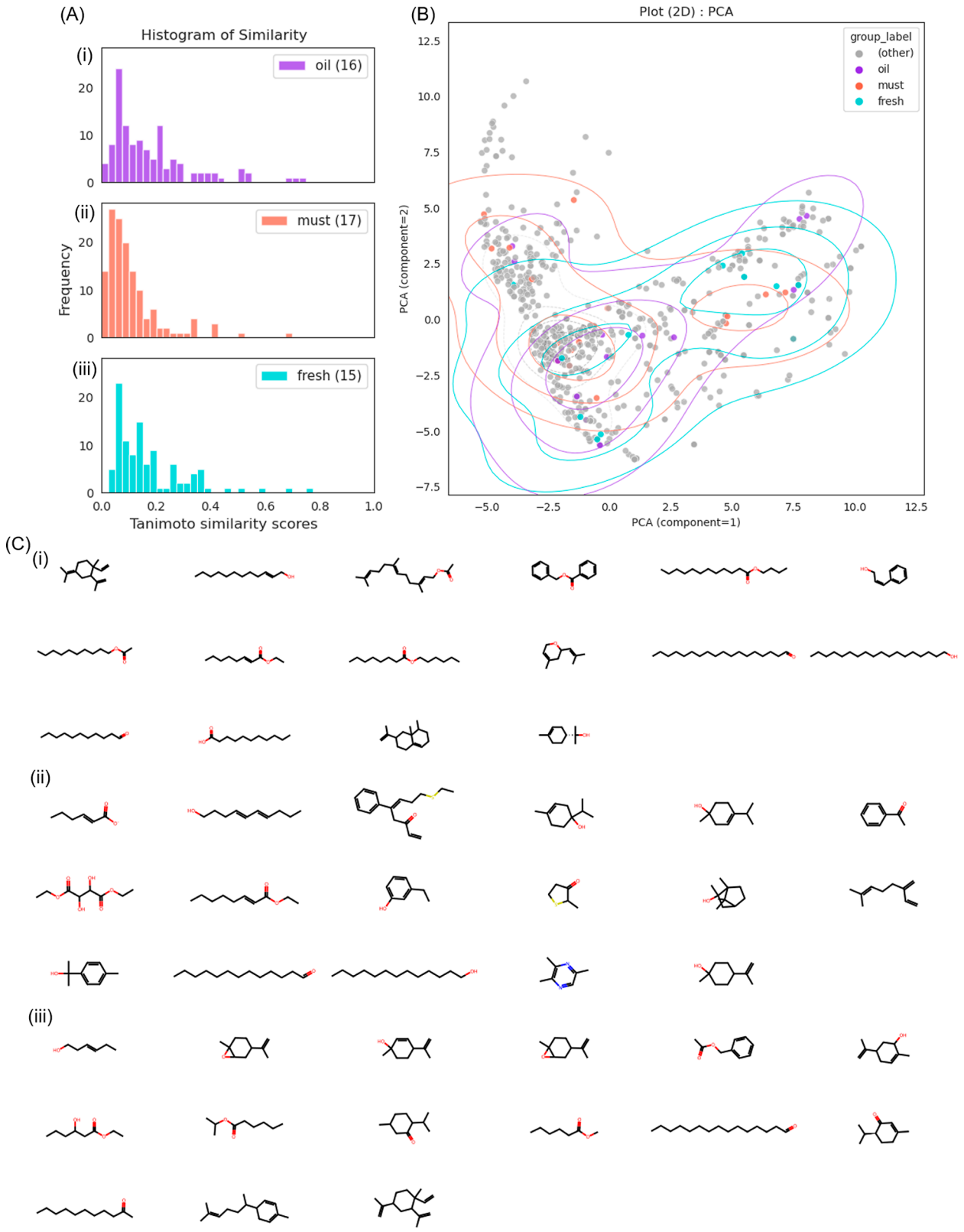


Figure 6. Structural diversity of the "low predictability odorants group"; (A) histogram of pairwise Tanimoto similarity scores in each odorants group and (B) mapping of principal component analyses based on RDKit fingerprints. (C) Molecules in the "low predictability odorants group" (i) "oil", (ii) "must", and (iii) "fresh".

odorants groups contain a relatively small number of chemicals: 15, 8, and 10, respectively. As remarked by "*1" in Table 2, their AUCs show relatively high values: 0.983, 0.958, and 0.997, respectively. As shown in Figure 5A and Table 2 (average of Tanimoto scores), the average of pairwise similarity values are relatively high: 0.321, 0.409, and 0.153, respectively, among these three odorants groups, meaning that these three odorants groups show narrow structural diversity

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

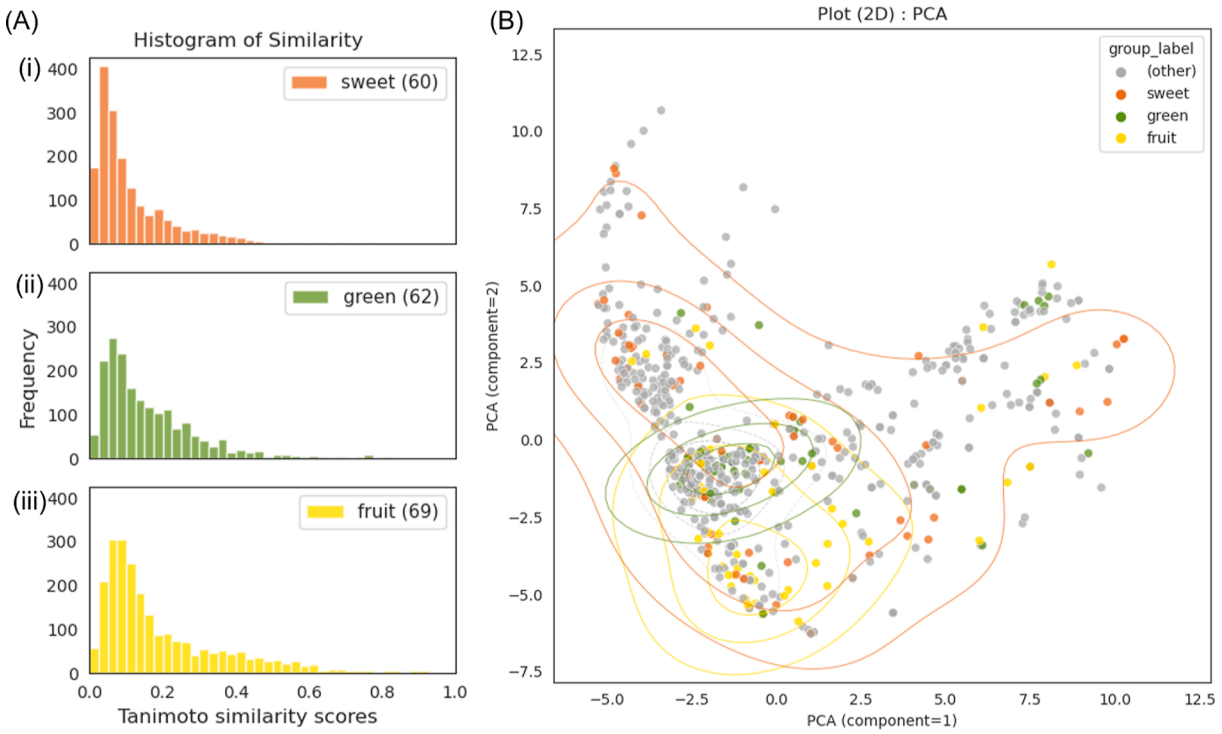


Figure 7. Structural diversity of "the odorants group with a large number of molecules"; (A) histogram of pairwise Tanimoto similarity scores in each odorants group and (B) mapping of principal component analyses based on RDKit fingerprints.

and consist of similar molecular structures. In Figure 5B, we plot the results of PCA mapping among the whole chemical space (shown in gray). It also shows that these three odorants groups have narrow structural diversity. In fact, they have relatively common building-blocks, functional groups, and/or skeletal molecular features, as shown in the molecules in Figure 5C. Especially "alkane" is an obvious example, where all molecules are evidently so similar (note the 13th row with "*1″ in Table 2, validation 1.00).

Although these three odor descriptors are used neither frequently nor popularly, their predictions worked relatively well. As shown in Table 2, for these descriptors, the regression shows a high potential for predictability (high validation values).

As shown in Figure 4, "cabbage (CA)" and "sulfur (SU)" are the outliers; these odorants groups have relatively wide structural diversity but still show "high predictability". Figure 5A also shows that "cabbage" has a relatively wider structural diversity compared to that of "cucumber" or "rose". The most probable explanation may be due to their chemical property; that is, the molecular structures of "cabbage" have a unique functional group (thioether, sulfide, thiocyanate, and/or thiophen-ring), as shown in the molecules of (iii) of Figure 5C. Thus, it shows a high AUC in odor descriptor prediction, as shown in Table 2. It also shows a narrow variance, as shown in Figure 5B. The "sulfur (SU)" is also an outlier. They obviously have the functional groups containing sulfur. The predictability is elevated, in spite of the large difference in their skeletal structure; most probably, this factor (the influence of atomic sulfur) is superior to the structural factor.

The descriptors "oil", "must", and "fresh" are the examples of "low predictability". These odorants groups contain a relatively small number of chemicals: 16, 17, and 15, respectively. As remarked in Table 2 by "*2", their AUCs in evaluation are small: 0.694, 0.429, and 0.691, respectively. Because the histogram bars in Figure 6A are left-skewed in some way, the pairwise similarity values are low, indicating a wide structural diversity in these odorants groups. Figure 6B shows the wide structural diversity of these odorants groups throughout the entire Flavornet chemical space. It is noteworthy that these odorants might have different building-blocks, functional groups, and/or skeletal molecular features, as shown in Figure 6C.

The result of regression analysis for these three "low predictability odorants groups" is potentially multimodal regression. These three odor descriptors are used in the odorant chemical space neither frequently nor popularly. Thus, it is possible to understand the difficulty to be predicted.

We now discuss the descriptor and predictability relation. The descriptors "sweet", "green", and "fruit" are examples of the "odorants group with a large number of molecules"; they contain a relatively large number of chemicals in the Flavornet database.

As shown in Table 2, for "sweet", "green", and "fruit", we found that the AUCs in training and in evaluation data have low and relatively medium values. Figure 7A shows that their pairwise similarity values in each odorants group are relatively low. Figure 7B also shows that these three odorants groups have wide structural diversity among the whole Flavornet chemical space. It suggests that these odorants groups might have relatively different building-blocks, functional groups, and/or skeletal molecular features.

When it comes to "the odorants group with a large number of molecules", the low AUC reflects the vagueness of these odor descriptors in regression analysis. These odor descriptors seemed to be frequently used for the odorant chemical space in human culture (natural language) in comparison with other descriptors. As a matter of fact, the descriptions cover a wide

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

range of meanings in natural language, and these odorants groups also contain wide structural diversity. These odorants groups have a relatively large number of chemicals, and as a result, the AUC turns out to be relatively small or medium (see also Figures 2 and 3 for the odor descriptor "fruit").

3.3. Additional Verification by Arctander and Goodscents. The human olfaction community has put considerable effort into making odorant-linked data sets such as Flavornet,[12] Arctander,[25] Goodscents,[26] and many more. In addition to the Flavornet that we mentioned now, we also carried out the analysis of the same regression study on Arctander, which contains 2751 molecules, and Goodscents, which contain 4565 molecules. The argument we obtained from these two data sets is consistent with the results that we have presented so far. The details and the discussion are shown in Supporting Information (S2−S4).

3.4. Resolution of Descriptors. In this study, we examined whether it is possible to clearly link the information space of odor descriptors with molecular structures. Previous researchers tried vector representations for each odorant and applied them to various regressions to solve the odor type prediction problem. Some related research studies are shown in Supporting Information (S6). The methodological accuracy of some modern models was reported by Gerkin; their AUC values were high compared to those of our simple regression in this study.[27] However, they faced other problems. Some researchers reported that some odorants groups have large structural diversity in them, which is sometime called "scaffold hopping".[28] Some researches introduced various methods and attempted to establish molecular structure-based descriptors. They have so far concluded that the prediction is very difficult due to the nature of the descriptors. Other research studies concluded that some odor descriptors are hard to predict due to large structural diversity within one odorants group.[2,4−8,27,29,30] This is a regression problem of two information spaces: the target variables are odor descriptors, whereas the explanatory variables are molecular structures. The predictability (accuracy) reflects the properties of two information spaces: complexity in odor descriptors and diversity in molecular structures. We have shown that the predictability of odor descriptors depends on the diversity of molecular structures within each odorants group.

The odor descriptors might have ambiguity. It might also have a kind of multimodality. Some researchers reported that odor descriptors have a vector representation based on NLP studies such as word2vec.[31−33] Kowalewski has also previously noted them in his paper.[34]

Is it possible to visualize and quantify the resolution of odor descriptors? Iatropoulos et al. introduce two new metrics: the olfactory association index (OAI, how strongly a word is associated with olfaction) and the olfactory specificity index (OSI, how specific a word is in its description of odors).[35] Ravia et al. introduced "olfactory metamers—pairs of non-overlapping molecular compositions that generated identical odor percepts".[36] Since both reports provided a degree of perceptual similarity, their concepts may be close to each other. The OSI and OAI are based on a text corpus study; on the other hand, "olfactory metamer" is based on a psychological experiment. They might be related to the resolution in the odor descriptors.

3.5. Hierarchical Definition of the Odor-Sensing Space. We previously defined the "odor-sensing space".[1] By using this definition, there are various "odor-sensing spaces" corresponding to each step in the olfactory system (Figure 1). In this study, we discussed a regression from one step in the chemical space to another step in the odor descriptor space. Thus, we faced difficulties with complications in the odor descriptor space. Therefore, the viewpoint of hierarchical odor-sensing space provides a clue.

The odor descriptors consist of natural language. Therefore, it is closely related to the odor-sensing space in the human brain. We can make further progress by considering the informatization of odors as a hierarchical process in the living body.

In parallel, the next subject to be carried out is an investigation of variable methods to extract the properties of molecules other than fingerprints, such as quantum chemical properties and vibrational modes. A molecular study using these approaches is now ongoing.

4. CONCLUSIONS

In this study, we report regression analysis and mapping based on the odorant chemical space. We examined whether it is possible to clearly link the information space of odor descriptors with molecular structures. The strong influence is traced between structural diversity and the predictability of the odor descriptor. We carried out the analysis by four fingerprints; the essential arguments are qualitatively conserved among the four fingerprints. In this investigation, we encountered difficulties related to the complexity included in the odor descriptor. The viewpoint of the hierarchical definition in the odor-sensing space will allow for further improvement in odor informatization.

■ ASSOCIATED CONTENT

Data Availability Statement

We used the RDkit[18] for the fingerprints (MACCS, ECFP, Avalon fingerprint, and RDKit fingerprint) and Tanimoto similarity scores of fingerprints. The regression methodology, multivariate analysis, and mapping are proprietary but not restricted to our program.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c02268.

Additional experimental details, discussions, and methods (PDF)

Exported Jupyter notebook of the regression analysis (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Yuki Harada − *Priority Organization for Innovation and Excellence Laboratory for Data Sciences, Kumamoto University, Kumamoto 860-8555, Japan;* orcid.org/0009-0000-4254-7803; Email: yharada@kumamoto-u.ac.jp

Authors

Shuichi Maeda − *Priority Organization for Innovation and Excellence Laboratory for Data Sciences, Kumamoto University, Kumamoto 860-8555, Japan*

Junwei Shen − *Priority Organization for Innovation and Excellence Laboratory for Data Sciences, Kumamoto University, Kumamoto 860-8555, Japan;* orcid.org/0000-0003-4223-6735

https://doi.org/10.1021/acsomega.4c02268
ACS Omega 2024, 9, 25054−25062

Taku Misonou − *Emeritus Professors of University of Yamanashi, Kofu 400-8510, Japan*
Hirokazu Hori − *Emeritus Professors of University of Yamanashi, Kofu 400-8510, Japan*
Shinichiro Nakamura − *Priority Organization for Innovation and Excellence Laboratory for Data Sciences, Kumamoto University, Kumamoto 860-8555, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c02268

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by Tateishi Science and Technology Foundation Research grant (A) no. 2221022.

■ REFERENCES

(1) Harada, Y. A Study for Odor Component Exploration with Multi-dimensional Data Analysis of Odor Sensing Spaces. Ph.D. Thesis, Tokyo Institute of Technology, 2016.
(2) Boelens, M.; Boelens, H.; Boelens, H. Some aspects of qualitative structure-odor relationships. *Perfum. Flavor.* **2003**, *28*, 36−45.
(3) Martinez-Mayorga, K.; Peppard, T. L.; Yongye, A. B.; Maggiora, G. M.; Medina-Franco, J. L. Flavor landscape: Towards a systematic characterization of a comprehensive flavor database. *Abstr. Pap. Am. Chem. Soc.* **2011**, *25* (10), 550−560.
(4) Keller, A.; Gerkin, R. C.; Guan, Y.; Dhurandhar, A.; Turu, G.; Szalai, B.; Mainland, J. D.; Ihara, Y.; Yu, C. W.; Wolfinger, R.; Vens, C.; et al. Predicting human olfactory perception from chemical features of odor molecules. *Science* **2017**, *355*, 820−826.
(5) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine learning for scent: Learning generalizable perceptual representations of small molecules. 2019, arXiv:1910.10685 arXiv preprint. https://doi.org/10.48550/arXiv.1910.10685.
(6) Sharma, A.; Kumar, R.; Ranjta, S.; Varadwaj, P. K. SMILES to smell: decoding the structure−odor relationship of chemical compounds using the deep neural network approach. *J. Chem. Inf. Model.* **2021**, *61*, 676−688.
(7) Licon, C. C.; Bosc, G.; Sabri, M.; Mantel, M.; Fournel, A.; Bushdid, C.; Golebiowski, J.; Robardet, C.; Plantevit, M.; Kaytoue, M.; Bensafi, M. Chemical features mining provides new descriptive structure-odor relationships. *PLoS Comput. Biol.* **2019**, *15*, No. e1006945.
(8) Lee, B. K.; Mayhew, E. J.; Sanchez-Lengeling, B.; Wei, J. N.; Qian, W. W.; Little, K.; Andres, M.; Nguyen, B. B.; Moloy, T.; Yasonik, J.; Parker, J. K.; et al. A principal odor map unifies diverse tasks in olfactory perception. *Science* **2023**, *381*, 999−1006.
(9) Spring, D. R. Diversity-oriented synthesis; a challenge for synthetic chemists. *Org. Biomol. Chem.* **2003**, *1*, 3867−3870.
(10) Burke, M. D.; Berger, E. M.; Schreiber, S. L. Generating diverse skeletons of small molecules combinatorially. *Science* **2003**, *302*, 613−618.
(11) Galloway, W. R. J. D.; Isidro-Llobet, A.; Spring, D. R. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **2010**, *1*, 80.
(12) Acree, T.; Arn, H. *Flavornet and human odor space*. https://www.flavornet.org/.
(13) Kaeppler, K.; Mueller, F. Odor classification: a review of factors influencing perception-based odor arrangements. *Chem. Senses* **2013**, *38*, 189−209.
(14) Nguyen, D. D.; Wei, G.-W. DG-GL: Differential geometry-based geometric learning of molecular datasets. *Int. J. Numer. Meth. Biomed. Eng.* **2019**, *35*, No. e3179.
(15) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273−1280.
(16) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208−220.
(17) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(18) *rdkit.org, RDKit: Open-source cheminformatics*. https://www.rdkit.org.
(19) James, C. A. *Daylight Theory Manual*, 2011. https://www.daylight.com/dayhtml/doc/theory/index.html.
(20) Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer-Verlag: Berlin, Heidelberg, 2006.
(21) *StratifiedKFold*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.
(22) Qian, W. W.; Wei, J. N.; Sanchez-Lengeling, B.; Lee, B. K.; Luo, Y.; Vlot, M.; Dechering, K.; Peng, J.; Gerkin, R. C.; Wiltschko, A. B. Metabolic activity organizes olfactory representations. *Elife* **2023**, *12*, No. e82502.
(23) Sorkun, M. C.; Mullaj, D.; Koelman, J. V. A.; Er, S. *ChemPlot, a Python Library for Chemical Space Visualization*, 2022.
(24) Gaytán-Hernández, D.; Chávez-Hernández, A. L.; López-López, E.; Miranda-Salas, J.; Saldívar-González, F. I.; Medina-Franco, J. L. Art driven by visual representations of chemical space. *J. Cheminf.* **2023**, *15*, 100.
(25) Arctander, S. *Perfume and Flavor Chemicals: Aroma Chemicals*; Allured Publishing Corporation, 1969.
(26) The Good Scents Company Information System. http://www.thegoodscentscompany.com/ (accessed on June 3, 2024).
(27) Gerkin, R. C. Parsing sage and rosemary in time: The machine learning race to crack olfactory perception. *Chem. Senses* **2021**, *46*, bjab020.
(28) Sun, H.; Tawa, G.; Wallqvist, A. Classification of scaffold-hopping approaches. *Drug discovery today* **2012**, *17*, 310−324.
(29) Rossiter, K. J. Structure- odor relationships. *Chem. Rev.* **1996**, *96*, 3201−3240.
(30) Keller, A.; Vosshall, L. B. Olfactory perception of chemically diverse molecules. *BMC Neurosci.* **2016**, *17*, 55.
(31) Nozaki, Y.; Nakamoto, T. Correction: Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. *PLoS One* **2018**, *13*, No. e0208962.
(32) 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*, No. e0234688.
(33) Shang, L.; Liu, C.; Tang, F.; Chen, B.; Liu, L.; Hayashi, K. Machine-Learning-Based Olfactometry: An Auxiliary System for Human Assessors in Olfactory Measurement. *bioRxiv* **2022**, 2022.04.20.488973.
(34) Kowalewski, J.; Huynh, B.; Ray, A. A system-wide understanding of the human olfactory percept chemical space. *Chem. Senses* **2021**, *46*, bjab007.
(35) 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.
(36) Ravia, A.; Snitz, K.; Honigstein, D.; Finkel, M.; Zirler, R.; Perl, O.; Secundo, L.; Laudamiel, C.; Harel, D.; Sobel, N. A measure of smell enables the creation of olfactory metamers. *Nature* **2020**, *588*, 118−123.