Fundamental Study on Chemical Data Analysis of Odors using Integrated Sensor Array and Machine Learning

チャイヤナ, ジラヨパット

https://hdl.handle.net/2324/4784659

出版情報：Kyushu University, 2021, 博士（工学）, 課程博士
バージョン：
権利関係：

Fundamental Study on Chemical Data Analysis of Odors using Integrated Sensor Array and Machine Learning

A DISSERTATION SUBMITTED TO

INTERDISCIPLINARY GRADUATE SCHOOL
OF ENGINEERING SCIENCE,
KYUSHU UNIVERSITY

IN PARTIAL FULFILLMENT OF REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN ENGINEERING

CHAIYANUT JIRAYUPAT

March 2022

Acknowledgments

This work was carried out at the Institute for Materials Chemistry and Engineering and the Department of Molecular and Material Sciences in Interdisciplinary Graduate School of Engineering Sciences at Kyushu University during 2018-2022. I would like to thank Japan International Cooperation Agency (JICA) for its financial support.

I would like to thank my supervisor, Prof. Takeshi Yanagida, for contributing me the opportunity to undertake my Ph.D. research program.

I would like to express my sincere appreciation to Prof. Kazuki Nagashima for his perspectives on scientific research. I also learned the importance of scientific and logical thinking from him.

I wish to thank the other laboratory staff, Prof. Tsunaki Takahashi, Prof. Takuro Hosomi, and Dr. Masaki Kanai, for their constructive suggestion and technical supports of various experiments for my research. I would like to take this opportunity to thank Prof. Guozhu Zhang, Prof. Zetao Zhu, Prof. Chen Wang, Dr. Yosuke Hanai, and Mr. Atsuo Nakao for their insightful advice and constructive discussions. Thank the secretary in our laboratory, Ms. Hiroko, Imaizumi, for her kind help in my daily life and study.

I would like to extend my sincere gratitude to our excellent collaborators, including Prof. Takao Yasui and Prof. Yoshinobu Baba from Nagoya University.

In addition, I thank all alumni and current members in our Laboratory: Dr. Benjarong Samransuksamer, Dr. Xixi Zhao, Mr. Kentaro Nakamura, Ms. Wenjun Li, Mr. Wang Qiuanli, Mr. Hiroshi Anzai, Mr. Masahiro Shimuzu, Mr. Satoru Shiraishi, Mr.

Mizuki Matsui, Mr. Hiroki Yamashita, Mr. Yuki Nagamatsu, Mr. Kentaro Nakabayashi, Mr. Yuya Akihiro, Mr. Sameh Okasha, Mr. Akihide Inoue,

Finally, I special thank my family and friends. Their understanding and encouragement greatly support me during my Ph.D. research.

Chaiyanut Jirayupat

Abstract

This thesis is devoted to a fundamental study on chemical data analysis of odors using integrated sensor array and machine learning. The chemical data analysis of odors plays an essential role in extracting crucial chemical information and monitoring physiology data for various applications in diverse fields such as the food industry, environmental monitoring, security surveillance, medicine, and healthcare which effectively improve the quality of the human lifestyle. However, although odor sensing-based integrated sensor arrays have demonstrated the feasibility of real-time monitoring of the above application, it is far from practical use, which will significantly restrict their further optimization and commercialization.

First, to extract the vast chemical information in odor, we present an automated method to identify multivariate chemo-/biomarker features of analytes in chromatography-mass spectrometry (MS) data by combining image processing and machine learning. Our approach allows us to comprehensively characterize the signals in MS data without the conventional peak picking process, which suffers from false peak detections. The feasibility of marker identification is successfully demonstrated in case studies of aroma odor and human breath on gas chromatography−mass spectrometry (GC−MS) even at the parts per billion level with a low error rate by comparison with the conventional method.

Second, by using the above approach, we demonstrate a preliminary study for the breath odor analysis and breath odor sensing-based individual authentication (fasting condition) using an integrated sensor array and machine learning. We successfully achieved a median accuracy of 96.4%. The impacts of several sensors (features) on the accuracy and reproducibility are demonstrated. However, the applicability of breath odor sensing in practical use needs to be justified.

Finally, we demonstrate a blood glucose prediction system by breath odor analysis. Multi-classification of blood glucose in a complex environment (fasting, drinking, exercise) can be achieved with high accuracy. Furthermore, recognizing different glucose spike patterns leads to blood glucose monitoring in daily life. Our findings in this study provide an essential foundation for a robust breath odor sensing-based integrated sensor array system in the predictable future.

CONTENTS

Acknowledgments i

Abstract iii

CONTENTS v

CHAPTER I GENERAL INTRODUCTION 2

1.1 Introduction 4

1.1.1 Why Do We Need Big Data in Chemistry? *4*

1.1.2 Attraction of Chemical Information in Odors *4*

1.1.3 Problem in Chemical Data Analysis in Odors. *5*

1.2 Framework of This Thesis.. 6

1.3 References... 8

CHAPTER II LITERATURE REVIEW 11

2.1 Introduction 13

2.2 Odor Analysis/Sensing Methods 13

2.2.1 Laboratory Instruments *14*

2.2.2 Integrated Sensor Array..*15*

2.3 Data Analysis based on Machine learning.....................................16

2.3.1 Linear Discriminant Analysis *17*

2.3.2 Random forest..*18*

2.3.3 Neural Network...*18*

2.3.4 Deep Neural Network..19

2.3.5 Cross-Validation 20

2.3.6 System Performance Evaluation..20

2.4 Application of Odor Analysis/Sensing 21

2.4.1 Food industry...21

2.4.2 Environment Evaluation...22

2.4.3 Healthcare..23

2.5 Summary of Literature Review..24

2.6 References...25

CHAPTER III Automated Identification of Chemo-/Biomarkers in Chromatography−Mass Spectrometry 30

3.1 Abstract 32

3.2 Introduction 33

3.3 Experimental Section...35

3.4 Results and Discussion..41

3.5 Conclusions..46

3.6 References..47

3.7 Supporting Information...50

CHAPTER IV Discrimination of Complex Odors with Gas Chromatography-Mass Spectrometry Data by Texture Image Analysis and Machine Learning 72

4.1 Abstract 74

4.2 Introduction 75

4.3 Experimental Section 76

4.4 Results and Discussion 80

4.5 Conclusion 88

4.6 References..89

CHAPTER V Human Breath Odor Based Individual Authentication by Integrated Sensor Array 92

5.1 Abstract 94

5.2 Introduction 94

5.3 Experimental Section 96

5.4 Results and Discussion 99

5.5 Conclusion 107

5.6 References..109

5.7 Supporting Information...112

CHPATER VI Activity-Tolerated Blood Glucose Monitoring in Breath Sensing 122

6.1 Abstract 124

6.2 Introduction 125

6.3 Experimental Section 126

6.4 Results and Discussion 131

6.5 Conclusions..143

6.6 References...144

CHAPTER VII OVERALL CONCLUSION AND OUTLOOK 146

LIST OF PUBLICATIONS...151

CHAPTER I

GENERAL INTRODUCTION

1.1 Introduction

1.1.1 Why Do We Need Big Data in Chemistry?

Nowadays, Big data has a strong impact in the diverse fields in chemistry for discovering new chemical information via artificial intelligence (AI)[1]. The advances in data computing[2], storing[3], and cloud technologies[4] have enhanced analysis capability to explore the crucial chemical mechanisms[5], predict a novel substance speedy, and so on. The concept of Big data can realize through the "3V" vision include 1) volume data: an increasing number of data to improve decision and prediction. 2) velocity: fast collecting data and processing to deal with large volumes of data. 3) variety: combining multiple data sources to generate a new ideal[1,6-9]. Based on these concepts, Big data in chemistry allows us to unlock new opportunities for real-time monitoring applications such as intelligent unit operations, intelligent processing, autonomous operations.

1.1.2 Attraction of Chemical Information in Odors

The chemical composition of an odor containing many volatile substances indicates a chemical change in substance, bio-chemical in living things, and participation in the biogeochemical cycle, which is responsible for the ecological cycle and the existence of living organisms.

Researchers worldwide utilize chemical information in odors for diverse applications in various fields. For example, it is used for 1) medicine and healthcare[10-12] to check dietary routines and conditions of health from exhaled breath or body odor for non-invasive and high-quality healthcare services. 2) environmental evaluation[13-15] to differentiate hazardous substances and detect harmful odors in living areas. 3) food monitoring[16-18] to determine the ripeness of fruits and freshness of meat in livestock. 4) security supervision[19,20] to detect explosive and flammable materials. 5) fragrance research[21,22] to predict the degeneration of cosmetic products and detergents.

Last decade, developing such a device, the so-called "electronic nose (e-Nose), has been rising interest." e-Nose is an electronic device capable of sensing, detecting, and recognizing many odors by integrating several artificial olfactory sensors. e-Nose allows us to detect the varieties of chemical information in odors quickly and in real-time. As a result, we can gain a large amount of acquisition data of odor. Such advantages of odor sensing can satisfy the chemical Big data concept (3V).

1.1.3 Problem in Chemical Data Analysis in Odors.

It Research and development work in odor analysis/sensing primarily reports the feasibility of odor analysis/sensing far from natural conditions. A complex mixture of odor produced by living things (i.e., breath odor and body odor) not only contains thousands of volatile compounds but is also highly diverse and has vast concentration ranges from parts per trillion (ppt) to parts per million (ppm) levels[23,24], which can be affected by the surrounding environment and daily-life activity. However, previous odor analysis/sensing research focused on specific chemical information in a complex mixture. They were tracking bio/chemo-marker related to their research goal. For example, many research work proposed breath acetone as a high correlated biomarker of diabetic diagnostics—increasing breath acetone results from intense gluconeogenesis in the liver, which generates glucose from non-carbohydrate[25-27]. Nevertheless, the liver also produces acetone during periods of caloric restriction of various scenarios such as fasting, intense exercise, and alcoholism. This evidence proves that humans have a sophisticated metabolism, and limited information is insufficient to monitor daily-life physiological information.

The testing might lead to the high error rate of physiology prediction through breath odor analysis/sensing and wrong scientific discovery. Moreover, most of such impractical tests showed in breath odor analysis work. Breath samples were regularly collected under fasting conditions (8-10 hours) far away from daily-life health monitoring and practical use[28]. Though the vast molecular sensing ability of odor sensing-based e-nose, why does research still focus on specific chemical information rather than a considerable number of chemical information? Inconsistency of data characteristics and data analysis point-of-view for odor sensing is a bottleneck for the development of daily-life health monitoring and practical use. The analytical platform to automatically analyze the enormous chemical data in odor is required.

1.2 Framework of This Thesis

This thesis consists of six chapters. We present all of the chapters as follows:

- Chapter I is an overview of the rationale for this research, including the importance of Big data in chemistry, the attraction of chemical information in odors, the current problems in chemical data analysis in odor based on sensors, and the framework of this thesis.
- Chapter II outlines the previous literature about the techniques of molecular gas detection, including some bulk laboratory instruments and portable metal oxide-based sensors. Again, we emphasized Metal oxide semiconductor nanowires-based sensors.
- Chapter III presents an automated method to identify multivariate chemo-/biomarker features of analytes in chromatography-mass spectrometry (MS) data by combining image processing and machine learning.

- Chapter IV demonstrates a preliminary study for the breath odor sensing-based individual authentication using an artificial olfactory sensor array.
- Chapter V demonstrates an activity-tolerated blood glucose monitoring by artificial intelligence-based ensemble feature analysis of breath sensing data.
- Chapter VI is an overall conclusion of this thesis and the perspective for possible future work.

1.3 References

1. Big data in chemistry (biomedcentral.com)
2. Ibrahim A, Targio H, Ibrar Y, Nor B A, Salimah M, Abdullah G, and Samee U K. Information system 2015,47, 98-115.
3. Mazumdar S, Seybold D, and Kritikos K. J Big Data 2019, 6, 15.
4. Aakash T, IJERT 2017, 5, 23.
5. Zachary J B, Xiang Y, Philippe Y A, Yanan Z, Steven P. W, and Qiongqiong Z, Journal of Chemical Information and Modeling 2021, 61, 3197-3212.
6. Big data - Wikipedia
7. Tetko I V, Engkvist O. J Cheminform 2020, 12, 74.
8. Ashiff K, Seetharaman A, Abhijit D. Authorea. July 07, 2021.
9. Isuru A U, Carina L G, Yoshiyuki Y, Michael A. T, Ahmet P, Brent R Y, Krist V G, Murat K, Christoph B. *Ind. Eng. Chem. Res.* 2020, 59,34, 15283-15297
10. Capuano, R., Santonico, M., Pennazza, G. Sci Rep 2015, 5, 16491.
11. Yang, H Y., Wang, Y C., Peng, H Y. Sci Rep 2021, 11, 103.
12. Lundström J N, Olsson M J. Vitam Horm. 2010, 83, 1-23.
13. Kota S, Ryo T, Takako S, Yuko K, Keiko K, Eri S, Kosuke M, Huynh T N, Gaku I, Koji T, Genki Y. ACS Sens 2018, 3, 8, 1592-1600.
14. Petra I, Bedr E A, Yohann N, Thierry T, Pascal T, and Denis T. ACS Sens 2021, 6, 11, 3824-3840.
15. Shivani D, Mehta B R, Tyagi A K, Kapil S. Sens.Inter 2021, 2, 2666-3511.
16. Takahashi Y, Nagayama S, Mori K. J Neurosci. 2004, 24, 40, 8690-8694.
17. Vera S, Ethan D E, Youchi M W, Constantin AV, Benjamin R M, Suchol S, Timothy M S. ACS Sens 2019, 4, 8, 2101-2108
18. Alphus D W, Manuela B, Sensors 2009, 9, 5099-5148

19. Jehuda Y. Anal.Chem 2003 75, 5, 98-105.

20. Ye Z, Liu Y, Li Q. Sens 2021, 21, 7620.

21. Eamsa-ard T, Swe M M, Seesaard T, Kerdcharoen., IEEE 7th Global Conference on Consumer Electronics (GCCE), 2018, 363-364.

22. Adak M F, Yumusak N. IEEE EUROCON 2017 -17th International Conference on Smart Technologies, 2017, 644-648.

23. Ahmed H J, Fahmida A, Sohini R, Yogeswaran U, Nezih P, and Shekhar B. ACS Sens 2018, 3, 7, 1246-1263.

24. Andreas T G, Sebastian A, Karsten K, Philipp A G, Arno S T, and Sotiris E P. ACS Sens 2019, 4, 2, 268-280.

25. Bovey F, Cros, J, Tuzson, B. Nutr & Dia 2018, 8, 50.

26. Zhennan W, Chuji W. J. Breath Res 2013, 7, 037109

27. Gus H, Shrinivas S, Martin G, Daniel L, Clare M, Rob P, Graham R, Grant A R, Katharine R O, 2021 J. Breath Res 2021, 15, 017101

28. Minh T C, Blake D R, Galassetti P R. Dia.Res.Clin.Pract 2012, 97, 2, 195-205.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Chemical data analysis of odor is essential and regularly used in medicine and healthcare[1-3], environmental evaluation[4-6], food monitoring[7-9], security supervision[10,11], fragrance research[12,13]. A broad category of machine learning techniques is applied, from biomarker evaluation/feature extraction based on untagged metabolomics/chemometrics to the classification and recognition of odor fingerprints extensively[14-16]. The applicability of machine learning for odor analysis shows promise in extracting a high level of sophisticated information from the large size of raw data[17], heterogeneous[18], and high-dimensional data sets[19]. We review machine learning as a chemical data tool for odor analysis with many potential applications, including basic concept and machine learning algorithm for biomarker evaluation/feature extraction from raw data based on standard gold analysis[20-22] (i.e., GCMS) and odor sensing via integrated sensor array[23-25]. We demonstrate fundamental concepts and the basis for applying these strategies to chemical data of odor. This chapter presents various gas molecular detection techniques, including the gold standard laboratory instruments as GC/MS and portable artificial olfactory sensors. Then, various data analyses based on machine learning, such as supervised learning[26], unsupervised learning[27], linear discriminant analysis[28], random forest[29], deep neural network[30], system performance evaluation[31], and so on, are demonstrated to understand their working principles.

2.2 Odor Analysis/Sensing Methods

As mentioned above in the introduction section, Researchers worldwide utilize odor analysis in diverse fields with different purposes. We here divide the main target of odor analysis into two issues: odor component identification and odor fingerprint recognition. First, we demonstrate odor analysis as a tool for explaining the mechanism or microbiological pattern based on metabolomics profile. Here, the researcher identified bio-/chemomarkers to clarify the original pathway. Second, the researcher often utilizes odor sensing to recognize or discriminate the complex odor (i.e., exhaled breath). The prominent character of such odor always contains several (a hundred to thousand compound) bio-/chemomarker which unclarified their original mechanism, microbiological pattern, or pathway[32-34]. So, the technology for odor analysis was selected depending on the research target.

2.2.1 Laboratory Instruments

Gas chromatography-mass spectrometry. Gas chromatography connected with mass spectrometry (GC-MS) has been a well-known gold standard tool for extracting qualitative chemical information in complex gaseous mixtures. Many researchers use GC-MS to identify the chemical composition in odor. Especially for sophisticated biological/chemical samples, GC-MS has a decisive advantage due to the detection limit of GC-MS reaching up to part per trillion (ppt level) and high reproducibility[35]. In addition, it allows us to separate individual compounds from a gas mixture's physical behavior in the mobile and stationary phases. It separates compounds by transferring the sample gas mixture into a chromatographic column that controls temperature via an inert gas[36]. Each compound drives at different speeds based on the interaction with the column material[37,38]. Compound with a strong interaction drives slower than compound with weak interaction. Finally, the mass spectrometer identifies the molecular fragment of an eluted compound. Nevertheless, it concerns weaknesses such as extended analysis time, the impossible for real-time analysis, and external or internal standards requirements[39].

2.2.2 Integrated Sensor Array

Metal-oxide sensors. Metal-oxide sensors are the most generally used sensor type in the odor sensing device because of their suitability for various gas species[40]. The application field of metal-oxide sensors-based odor sensing is mainly associated with quality management, monitoring function, pollutant, contamination, and decomposition of food[41]. In addition, these sensors can work at high temperatures while requiring high energy consumption[42]. We categorize metal-oxide sensors into two main groups according to their electrical properties: n-type and p-type[43]. N-type metal-oxide sensors working principle relates to the reactions between the oxygen molecules and their surface. Free electrons on the metal oxide surface are trapped, resulting in potential barriers between grain boundaries that inhibit carrier mobility and generate extensive resistance regions. In contrast, P-type sensors react to the oxidizing molecules by producing holes and removing electrons. Their typical surface interaction and oxygen absorption/desorption enormously raise the sensor's performance while enhancing the sensing recovery, boosting the molecular selectivity, and decreasing the humidity effect. Such an advantage, these sensors are often selected in many odor sensing applications[44].

Conducting polymer sensors. Conducting polymers are proper candidate sensors compared with metal oxide sensors. Therefore, they are selected as a reliable sensor type for various odors sensing applications such as medical, pharmaceutical, food, and cosmetic industries. Sensors' materials are low cost, and the fabrication process is easy to perform on a large scale. Moreover, conducting polymers sensors respond and recovery fast to odorants[43]. In addition, the researcher can optimize the selectivity of their sensors by selecting several kinds of polymers such as polypyrrole, polyaniline, polythiophene, and others[45]. The resistance in the sensor changes due to an interaction with an analyte, resulting in the detection of odor.

2.3 Data Analysis based on Machine learning

Machine learning (ML) involves diverse mathematic theories such as probability, statistics, approximation. The machine learning algorithms can assemble a mathematical model using training data to make decisions themselves[46]. The performance of machine learning is most high-light in diverse scientific and engineering research fields: image[47], sound[48], language recognition[49] and multimedia[50]. Recently, machine learning has received significant attention, especially for analytical chemistry, biology, material science, biometric authentication, and robotics. The type of learning required in these tasks is detecting or classifying patterns, extracting features or expressions of raw data. We can demonstrate the concept of machine learning approaches into three types, relying on the character of signal and response functions to the learning system, including supervised[51], unsupervised[52], and reinforcement learning[53]. Here we only demonstrate supervised and unsupervised learning, which researcher wildly used for chemical data analysis of odor[54]. Conversely, reinforcement learning is applicable for an automated driving vehicle or playing a game against an opponent[55], which is not demonstrated in this thesis.

Supervised Learning: the main target in supervised learning is to learn a model from labeled training data[56]. Subsequently, a model can predict using unseen data or validation and test data. The training set for supervised learning must contain features as inputs data and answer or label as outputs data. Supervised learning models allow us for regression and classification models: naive Bayes classification can conduct only classification. While Linear and Logistic analysis, k-Nearest neighbor, Random forest can manipulate both classification and regression analyses.

Unsupervised Learning: On the contrary, unsupervised learning does not require an artificially labeled training set[57]. The researcher always utilizes standard unsupervised learning algorithms for anomaly detection, association, autoencoders, and data clustering. In addition, there is also semisupervised learning, which utilizes the synergetic between supervised and unsupervised learning. The researcher mainly used this method to extract significant features from medical images or spectroscopy mapping from complex data. We next briefly demonstrate a machine-learning algorithm widely used to analyze the chemical data of odor[58].

2.3.1 Linear Discriminant Analysis

Data scientists generally used linear discriminant analysis for supervised classification problems. It is an easy and effective linear classification model for binary problems; however, it can be applied for multiple classifications. The linear discriminant analysis allows data preprocessing to decrease the number of features, significantly reducing the analysis time[59,60]. In addition, linear discriminant analysis conduct simplifying hypotheses of the analyzed data. Finally, the method effectively deals with successive quantities of independent variables with unique observations[60]. We briefly explain the linear discriminant analysis by considering a generic binary classification problem: the model uses both the X and Y axes to generate a new axis. Then, the data is projected on the new axis to maximize the separation of the two classes and reduce information from two dimensions into one dimension. Finally, the considering two criteria are used to generate this new axis: 1) the distance between the two classes must be maximized and 2) the variation within each class must be minimized[61]. By reducing the number of significant data dimensions, the data scientist usually uses linear discriminant analysis to visualize the high-dimension problem. However, linear discriminant analysis has many advantages, as mentioned above. Nevertheless, it still has several limitations, such as the linear function cannot deal with more complex and multi-dimensional data.

2.3.2 Random forest

Random forest , known as, Random decision forest is categorized as unsupervised learning and a tree-based learning method that can solve classification and regression problems[62]. The fundamental concept behind the random forest model consists of various classifiers in one model. The classifiers or decision trees in a random forest model use the bootstrap aggregating technique to boost the performance of classification and regression compared with a single decision tree, resulting in handling enormous data sets with higher dimensionality[63]. Finally, we use all decision trees to calculate prediction accuracy by the mean value of the provided accuracy. Although increasing the number of trees in the forest can boost the model's prediction accuracy, the overfitting problem might be rising. In addition, the random forest model cannot further predict the over-range of training data for a regression problem[64].

2.3.3 Neural Network

Data scientists design neural networks to mimic the human brain functions to recognize the patterns loosely. For example, neural networks can interpret sensory data through machine perception and cluster or classify unlabeled and labeled data. A standard supervised learning neural network consists of 3 layers: input, hidden, and output. The training set contains values of the feature inputs: x and the answer outputs: y[65]. The neural network consists of a single layer of multiple neurons, so-called hidden layers[66]. The neuron is a mathematical operation that classifies information according to a characteristic structure. The network of neurons handles a substantial similarity to statistical analysis like multiple-linear regression. However, the hidden layer implies that the proper values for these nodes are not considered during learning. Instead, it refers to the values x of each layer carried on to the successive layers and finally generates the sets of activations[67]. In order to calculate the outputs of the model, the generic function such as sigmoid[68], hyperbolic tangent (tanh)[69], or the rectified linear unit (ReLU)[70] is frequently utilized as an activation function.

2.3.4 Deep Neural Network

The machine learning expertise has recently utilized a deep neural network to analyze scientific and mathematic information rather than the single hidden layer neural network. The meaning of "deep" in the deep neural network refers to the depth of hidden layers in a neural network. We can simplify the deep neural network as a hierarchical network of multiple neurons. The primary deep neuron network passes input signals in the first layer through neurons to other neurons that learn with feedback, so-called feed-forward. Finally, the output signals present a prediction result as "Yes or No" or show in probability ("0 or 1")[71].

Each layer can have from a single to multiple neurons, and each of them can generate a unique function such as an activation function, dropout, and others. For example, the activation function simulates the signal for transferring the signal to the subsequently connected neurons. If the transferred signal has a value greater than a threshold value, the output is transferred else rejected. Finally, the weight is evaluated between two successive neurons. Weight refers to the significance of input and output for the next neuron. The initial weights are randomly generated and recalculated iteratively during the training and iteration until the model is optimized.

The deep neural network also relies on backpropagation. The backpropagation allows the information to flow backward from the cost to compute the gradient more efficiently. Although deep neural networks allow a high performance with automatic feature extraction, the model requires an enormous data set to suppress the overfitting[72].

2.3.5 Cross-Validation

Cross-validation is a resampling strategy employed to evaluate the reproducibility and stabilize the performance of machine learning models when dealing with unseen data, especially for the limited data number. When conducting (supervised) machine learning, data scientists utilize k cross-validation to avoid common overfitting problems. The basic approach is as follows: first, the training data is divided into equal-sized "k." Then, each unique subgroup is separated as a training and testing data set. Next, tuning parameters of the model on the training data set and evaluating performance on the test set. Finally, the k-fold statistical estimation value is presented as a mean value of all optimized models. Moreover, there is also various k cross-validation as the nested cross-validation method, Leave-p-out cross-validation, Monte Carlo cross-validation, and other[73].

2.3.6 System Performance Evaluation

A confusion matrix is utilized to demonstrate the classification efficiency. In addition, the confusion matrix is beneficial to evaluate system performance like accuracy, precision, sensitivity, specificity, and others. We briefly explained how to interpret the confusion matrix on the binary classification. The two possible predicted categories are "correct and incorrect." The correct prediction: true-positive (TP) and true-negative (TN), A true-positive result from the model correctly predicts the "positive class." Likewise, a true-negative result is the corrected prediction of the "negative class." For the incorrect prediction: false-positive (FP) and false-negative (FN). A false-positive (type I error) result from the model incorrectly predicts the "positive class." Likewise, a false-negative (type II error) result is the uncorrected prediction of the "negative class."[74]

Accuracy is expressed as the correctly overall predictions and calculated as

$$\text{Accuracy} = TP + TN / TP + TN + FN + FP. \ (1)$$

Precision is independent with accuracy. Precision can refer to repeatability and reproducibility, which calculates as

$$\text{Precision} = TP / TP + FP. \quad (2)$$

Sensitivity refers to the proportion of correctly predicted positive results and calculates as

$$\text{Sensitivity} = TP / TP + FN. \quad (3)$$

Specificity refers to the proportion of correctly classified negative results, calculates as

$$\text{Specificity} = TN / TN + FP. \quad (4)$$

2.4 Application of Odor Analysis/Sensing

2.4.1 Food industry

The most powerful odor sensing application within farming has been food production. As a result, there has been considerable interest in using odor sensing systems to analyze and monitor chemical information in food odor for various applications in the food industry.

Odor sensing systems allow automated systems to rapidly predict food products to maintain product quality, safety, and nutrition based on chemical compound sensing. The odor sensing system is applicable for a quality guarantee of natural and manufactured products, monitoring fermentation, mixing, seasoning, and packaging process, determining day-to-day freshness and aging in livestock, and evaluating ripening and stage. Such capability provides that the final products are delivered to the customer with good-agreed quality in commercial markets. h.

2.4.2 Environment Evaluation

The emission of unpleasant odors from industry has increased significantly, damaging people's health worldwide. Primarily, industrial plants close to the residence place directly affect the rising number of unhealthy people. For example, an unpleasant odor might contain hundreds to thousands of chemical compounds, a solid toxic property such as Arsine, Hydrogen Sulfide, and Hydrogen Cyanide. Such toxic gas might affect the human body in short to long term. The effect leads to severe diseases and symptoms like lung cancer, skin cancer, unconsciousness, memory loss, respiratory infection, instant death, etc.

Accordingly, monitoring emission odor from the industrial area has become increasingly important to control environmental quality. For this purpose, an environmental impact assessment of odors is carried out, and appropriate measures are taken to decrease the emission of odors. For several industrial factories located in a relative area, researchers used odor analysis methods to identify the sources of emissions of odorous substances. This issue allows for identifying the factors responsible for the most incredible odor nuisance and implementing the procedures used to reduce the number of odorous substances released into the environment. For this purpose, olfactometry methods. When using the method of olfactometry, it is possible to determine the level of concentrations of odorous substances and the factor of odor emissions. Furthermore, using the odor analysis methods, it is also possible to predict the correlation between the concentration of the odorous substance and the weather conditions characteristic of the season.

2.4.3 Healthcare

Currently, breath odor becomes the non-invasive way for diagnosis and health monitoring. Many researchers propose the feasibility of utilizing breath analysis to predict fatal diseases and especially for respiratory diseases such as lung cancer, asthma, and others. They discovered the connection between a disease and breath components. Researchers worldwide evaluate that breath odor analysis has a high potential as the next generation of disease diagnostic and health monitoring in daily life. Clinical diagnostic uses costly analytical tools by blood testing, urine analysis, endoscopy, biopsy, imaging, and others. Due to complex protocols, these methods require a time-consuming sample collection procedure, experienced operators, and professional analysis. The diagnosis is limited to a specific place such as a hospital or clinic. In contrast, operators and users can efficiently diagnose disease using breath odor analysis. In addition, it allows us to monitor health conditions in real-time and everywhere, which effectively prevents fatal disease. It is an entirely non-invasive technique that allows the development a user-friendly, convenient, and intuitive diagnostic platform. The sample collection can be archived easier than blood, serum, urine, and other methods. Moreover, we do not need to be concerned about bio-hazardous specimens within regulations.

Diagnosis is based on scientific evidence of the physiological phenomenon; breath contains the Volatile organic compound (VOCs) or the biomarkers that can be recognized and discriminated against by different diseases. Unbalanced or abnormal metabolism in the human body affects change in concentration variation of produced VOCs, then molecule exchange occurs in the alveoli and is finally released from the body through the lungs during the respiratory process. The various metabolic and pathological disorders could be observed. According to the results of recent breath analysis, such as diabetes, cardiovascular disease (CVD), bacterial infertility, asthma, cancer, inflammatory disease, Alzheimer’s disease, and other diseases.

2.5 Summary of Literature Review

This chapter reviews the literature on odor analysis and sensing using different methods (GC-MS, PTR-MS, metal oxide sensor, and conducting polymer sensors). The bulk and expensive GC/MS, PTR-MS instruments undoubtedly could give the precise analysis for chemical information. The integrated sensor array was then mainly discussed from the aspects of the sensing mechanism and their applications. Finally, data Analysis based on Machine learning was widely used to analyze the chemical and biochemical information. Another challenge in applying machine learning in chemical odor research is the lack of data for reliable model development. Since the experiments in this area are expensive (i.e., human biochemical samples), most research only focuses on limited available data sets and areas. So we think that extracting a high significant feature plays a crucial role in overcoming this problem.

2.6 References

1. Capuano, R., Santonico, M., Pennazza, G. Sci Rep 2015, 5, 16491.
2. Yang, H Y., Wang, Y C., Peng, H Y. Sci Rep 2021, 11, 103.
3. Lundström J N, Olsson M J. Vitam Horm. 2010, 83, 1-23.
4. Kota S, Ryo T, Takako S, Yuko K, Keiko K, Eri S, Kosuke M, Huynh T N, Gaku I, Koji T, Genki Y. ACS Sens 2018, 3, 8, 1592-1600.
5. Petra I, Bedr E A, Yohann N, Thierry T, Pascal T, and Denis T. ACS Sens 2021, 6, 11, 3824-3840.
6. Shivani D, Mehta B R, Tyagi A K, Kapil S. Sens.Inter 2021, 2, 2666-3511.
7. Takahashi Y, Nagayama S, Mori K. J Neurosci. 2004, 24, 40, 8690-8694.
8. Vera S, Ethan D E, Youchi M W, Constantin AV, Benjamin R M, Suchol S, Timothy M S. ACS Sens 2019, 4, 8, 2101-2108
9. Alphus D W, Manuela B, Sensors 2009, 9, 5099-5148
10. Jehuda Y. Anal.Chem 2003 75, 5, 98-105.
11. Ye Z, Liu Y, Li Q. Sens 2021, 21, 7620.
12. Lundström J N, Olsson M J. Vitam Horm. 2010, 83, 1-23.
13. Kota S, Ryo T, Takako S, Yuko K, Keiko K, Eri S, Kosuke M, Huynh T N, Gaku I, Koji T, Genki Y. ACS Sens 2018, 3, 8, 1592-1600.
14. Drupad K. T, Eleanor S, Yun X, Depanjan S, Caitlin W D, Camilla L, Phine B, Joy M, Monty S, Tilo K, Royston G, Perdita B, ACS Central Science 2019, 5, 4, 599-606.
15. Wu J, Zan X, Gao L, JMIR Med Inform. 2019, 7, 3, e13476.
16. Koga N, Hosomi T, Zwama M, Jirayupat C, Yanagida T, Nishino K, Yamasaki S. Front Microbiol 2020, 16, 11, 581571.
17. L'Heureux A, Grolinger K, Elyamany F H, Capretz M A M, in IEEE Access 2017, 5, 7776-7797.

18. Qiu, J., Wu, Q., Ding, G. J. Adv. Signal Process. 2016, 67
19. Huang B G, Zhou H, Ding X, Zhang R, Man, and Cybernetics, Part B (Cybernetics) 2012, 42, 2, 513-529.
20. Xinhui L, Wei K, Weimin S, Qi S,Chem Intel Lab Sys 2016, 155, 145-150.
21. Smolinska A, Hauschild A-Ch, Fijten R R R, Dallinga J W, Baumbach J, Schooten F J, J. Breath Res 2014, 8, 027105.
22. Ana M J, Luis C, Opi Food Sci 2021, 37, 76-82.
23. N U Hasan, N Ejaz, W Ejaz, H S Kim, Sensors 2012, 12, 15542-15557.
24. Chuanjun L, Hitoshi M, Kenshi H, Sen & Act B Chem 2022, 351, 130960,
25. Genva M, Kenne K T, Deleu M, Lins L, Fauconnier M L. Int. J. Mol. Sci. 2019, 20, 3018.
26. Nasteski V. Horizons 2017, 4, 51-62.
27. Ghahramani Z. Unsupervised Learning. In: Adv Lec on ML 2003. Springer 2004, 3176, Berlin, Heidelberg.
28. Xanthopoulos P, Pardalos P M, Trafalis T B. Linear Discriminant Analysis. In: Ro Data Min 2013. Springer, New York, NY.
29. Qi Y. Random Forest for Bioinformatics. In: Zhang C., Ma Y. (eds) En ML 2012. Springer, Boston, MA.
30. Albawi S, Mohammed T A, Al-Zawi S, Int Con Eng Tech 2017,1-6.
31. Tarca A L, Carey V J, Chen X, Romero R, Drăghici S, PLoS Comput Biol 2007, 3, 6, e116.
32. Giannoukos S, Agapiou A, Taylor S, J Breath Res 2018 12, 2, 027106.
33. Mazzatenta A, Pokorski M, Di Giulio C, Physiol Rep 2021, 9, 18, e15034.
34. José A. S, Ralf Z, Chahan Y, Anal Chem 2014, 86, 23, 11696-11704
35. Beale, D.J., Pinu, F.R., Kouremenos, K.A. et al. Metabolomics 2018, 14, 152.

36. Gas Chromatography. (2020, August 16). https://chem.libretexts.org/@go/page/301
37. Hussain S Z, Khushnuma M. “GC-MS: Principle, Technique and its application in Food Science 2014.
38. Xiao J F, Zhou B, Ressom H W, Trends Analyt Chem. 2012, 32, 1-14.
39. Prazeller P, Palmer PT, Boscaini E, Jobson T, Alexander M, Rapid Commun Mass Spec. 2003, 17, 14, 1593-1599.
40. Wang C, Yin L, Zhang L, Xiang D, Gao R, Sensors 2010, 10, 2088-2106.
41. Berna, A. Sensors 2010, 10, 3882-3910.
42. Yamazoe N, Kurokawa Y, Seiyama T, Sensors and Actuators 1993, 4, 283-289,
43. Arshak K, Moore E, Lyons G M, Harris J, Clifford S. Sensor Review 2004, 24, 2, 181-198.
44. Hyo-Joong K, Jong-Heun L, Sensors and Actuators B: Chemical 2014, 192, 607-627.
45. Bai, H.; Shi, G. Sensors 2007, 7, 267-307.
46. H. Wang, C. Ma and L. Zhou, International Conference on Information Engineering and Computer Science, 2009, 1-4,
47. Wäldchen J, Mäder P, Methods Ecol Evol 2018, 9, 2216– 2225.
48. da Silva B, W. Happi A, Braeken A, Touhafi A, Appl. Sci 2019, 9, 3885.
49. Agarwal A, Thakur M K, Sixth Int Con Cont Com 2013, 181-185,
50. Saeed F, Paul A, Hong W H, Multimed Tools Appl 2020, 79, 16201–16217.
51. Tammy J, Jaimie L G, Anthony J R, Behavior Therapy 2020, 51, 5, 675-687.
52. Sinaga K P, Yang M, in IEEE Access 2020, 8, 80716-80727.
53. Akbar T, Amirhessam T, Wolfgang B, Amir H G, ACM Com Sur 2020, 54, 8, 1-35.
54. Gutierrez-Osuna R, in IEEE Sensors Journal 2002, 2, 3, 189-202.
55. Du W, Ding S. Com Sci 2019, 46, 8, 1-8.
56. Ang J C, Mirzal A, Haron H, Hamed H N A, Comp Bio Bioform 2016, 13, 5, 971-989.

57. J. Latif, C. Xiao, A. Imran, S. Tu, 2nd Int Con Com, Math Eng Tech 2019,1-5.

58. Jörn L, Dario K, Thomas H, Chem Sen 2019, 44, 1, 11–22.

59. Tharwat A, Linear Discriminant Analysis: A Detailed Tutorial 2017, 169 – 190.

60. J. Wen, Tran Cir Sys Vdo Tech 2019, 29, 2, 390-403.

61. Ioffe S. Probabilistic Linear Discriminant Analysis. In: Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, 3954. Springer, Berlin, Heidelberg.

62. Biau, G., Scornet, E. TEST 2016, 25, 197–227.

63. Mariana B, Lucian Dr, J Photo Rem Sens 2016,114, 24-31.

64. Paul A, Mukherjee D P, Das P, Gangopadhyay A, Chintha A R, Kundu S, Trans Img Pro 2018, 27, 8, 4012-4024.

65. https://www.investopedia.com/terms/n/neuralnetwork.asp

66. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

67. https://en.wikipedia.org/wiki/Neural_network

68. Sridhar Narayan, Info Sci 1997, 99, 2, 69-82.

69. Lau M M, HannL, Biomed Eng Sci 2018, 686-690.

70. Brownlee J, Machine Learning Mastery 2019.

71. LeCun, Y., Bengio, Y. Hinton, G. Nat 2015, 521, 436–444.

72. https://www.rfwireless-world.com/Terminology/Advantages-and-Disadvantages-of-Deep-Learning.html

73. https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

74. Ajay K, Deri C, Feras A B, Data Demo 2020,83-106.

CHAPTER III

Automated Identification of Chemo-/Biomarkers in Chromatography−Mass Spectrometry

3.1 Abstract

We present a method named *NPFimg*, which automatically identifies multivariate chemo-/biomarker features of analytes in chromatography–mass spectrometry (MS) data by combining image processing and machine learning. *NPFimg* processes a two-dimensional MS map (*m*/*z* vs retention time) to discriminate analytes and identify and visualize the marker features. Our approach allows us to comprehensively characterize the signals in MS data without the conventional peak picking process, which suffers from false peak detections. The feasibility of marker identification is successfully demonstrated in case studies of aroma odor and human breath on gas chromatography–mass spectrometry (GC–MS) even at the parts per billion level. Comparison with the widely used *XCMS* shows the excellent reliability of *NPFimg*, in that it has lower error rates of signal acquisition and marker identification. In addition, we show the potential applicability of *NPFimg* to the untargeted metabolomics of human breath. While this study shows the limited applications, *NPFimg* is potentially applicable to data processing in diverse metabolomics/chemometrics using GC–MS and liquid chromatography–MS. *NPFimg* is available as open source on GitHub (http://github.com/poomcj/NPFimg) under the MIT license.

Keywords: *Additives, Molecules, Aroma, Mathematical methods, Biomarkers.*

3.2 Introduction

Untargeted metabolomics/chemometrics have gained much attention in diverse fields including pathology, microbiology, pharmacology, food industry, environmental evaluation, and healthcare.[1−11] In these studies, mass spectrometry (MS) coupled with gas chromatography (GC−MS) or liquid chromatography (LC−MS) is widely used, and the features of chemo-/biomarker molecules in analytes are identified from their mass chromatogram data. The goal of untargeted metabolomics/chemometrics is to comprehensively characterize the chemo-/biomarker molecules in analytes. Analytes in biology and healthcare fields usually consist of a huge number of chemical components with various concentrations. Also, the reliable chemo-/biomarker characterization needs the examination of many samples. In this respect, a reliable sample characterization technique and a data analysis method for automated marker identification are strongly desired. To date, most efforts have been devoted to sufficiently extract the marker features in MS data. Recent development in MS technology allows for signal detection of important molecules in analytes at an ultratrace level.[12−16] This development of analytical hardware substantially expands the applicable field of research in metabolomics/chemometrics. On the other hand, the data processing in raw MS data remains a challenging issue. In general, there are two major tasks in the raw GC− or LC−MS data processing including peak picking and subsequent pairwise peak list comparison.[17−25] Various software resources including XCMS, [17,18] MZmine, [19,20] TracMass, [21] KPIC, [22] and others[23−25] have been developed to perform these tasks. However, they often suffer from insufficient peak picking performance. The peak picking algorithm in the abovementioned software resources is based on a binary (“peak” or “noise)” output method, in which the chromatographic peaks with a satisfactory shape (e.g., Gaussian), signal-to-noise ratio, and peak width are extracted by a thresholding approach. Such a thresholding approach usually causes many false positive/false negative peak detections. For example, less restricted threshold setting increases the number of false positive peaks, while the larger number of features can be extracted. Contrary, highly restricted threshold setting yields false negative peaks, while the fidelity of extracted peaks can be improved. These false detections in the peak picking process leads to wrong scientific discoveries and interferes with the interpretation of the correct ones. To solve the problem in the peak picking process, various machine learning-assisted techniques have recently been developed, which are based on support vector machine,[26] Bayesian optimization,[27,28] deep learning,[29,30] and others.[31] The former one automates the optimization of threshold parameter settings in the conventional software, for example, XCMS, and the latter two improve the peak/noise discrimination performance via recognizing the peak shape in computer vision. Such machine learning-assisted techniques successfully improved the peak picking performance compared with conventional software resources. However, these methods are complex and time-consuming because peak shape needs to be trained in advance by creating an original database. In addition, a peak/noise discrimination for trace-level molecules is a challenging issue because the shape recognition of a peak of low signal-to-noise ratio is difficult. Especially, the automated characterization of trace-level molecules in complex analytes (e.g., human breath), in which both high concentration and low concentration molecules coexist, is difficult. Thus, an automated data processing tool, capable of characterization of numerous molecules including trace-level ones, is strongly desired in untargeted metabolomics/chemometrics. In this work, we present a method named NPFimg, which automatically identifies multivariate chemo-/biomarker features of analytes in chromatography−MS data by combining image processing and machine learning. NPFimg processes a two-dimensional (2D) MS map to comprehensively characterize MS data, discriminate analytes, and identify and visualize marker features without the conventional peak picking process. The feasibility of chemo-/biomarker characterization is successfully demonstrated in case studies of aroma odors and human breath at various molecular concentration ranges [down to parts per billion (ppb) level]. The reliability of NPFimg is discussed by comparing it with the widely used XCMS. Furthermore, the applicability of NPFimg to untargeted metabolomics is examined via the human breath samples.

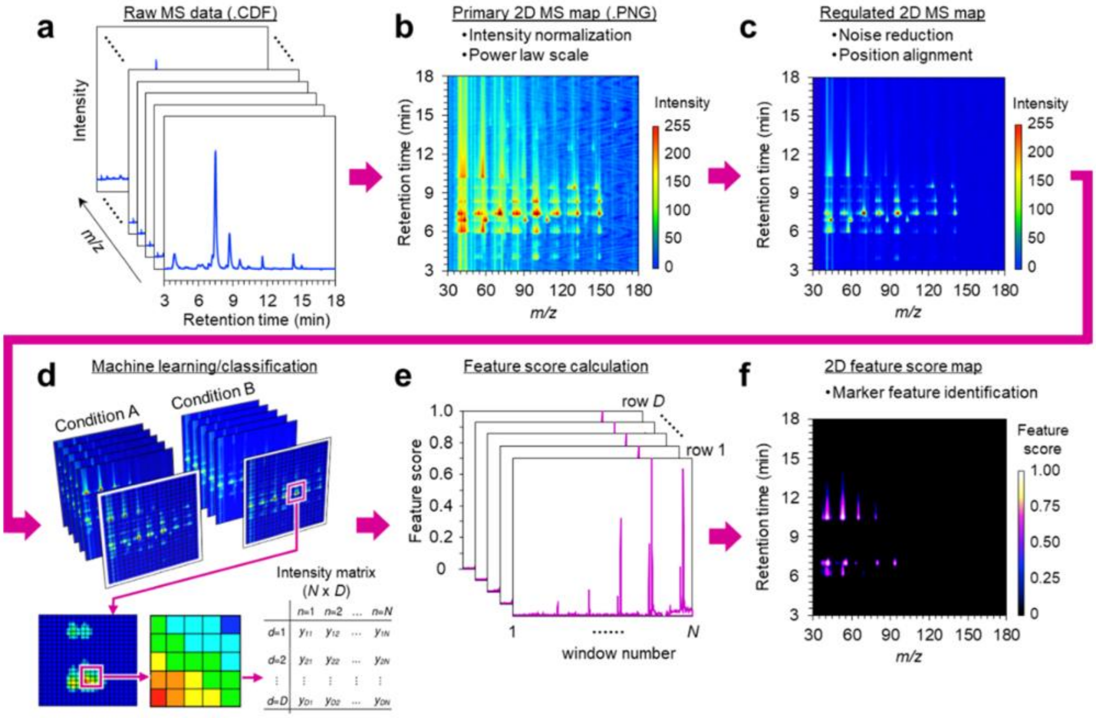
3.3 Experimental Section

Sample Preparation. We evaluated the performance of NPFimg to identify the chemo-/biomarker features in analytes by using aroma odor samples and human breath samples. In this study, the samples containing chemo-/biomarker molecules were prepared by adding the marker molecules to the original aroma odor/breath samples. For the aroma odor samples, we employed three types of commercial aroma oil including bergamot organic essential oil (aroma#1, Neal's Yard Remedies Inc.), lavender essential oil (aroma#2, Neal's Yard Remedies Inc.), and blended essential oil (aroma#3, Ryohin Keikaku Co., Ltd.). To collect the aroma odors, 50 μL of the aroma oil was first taken in a 20 mL vial bottle and it was left for 10 min at room temperature for fulfilling the vial bottle separated ports; one port was connected to an adsorbent-filled tube (Packed Liner with Tenax GR, mesh 80/100 #2414− 1021, GL Science Inc.), and the other port was connected to a nitrogen gas cylinder (99.997% pure). The other side of the adsorbent-filled tube was connected to an automatic air sampling pump (GSP-400FT, GASTEC Corp.). Then, 100 mL of the aroma odor was transferred from the headspace of the vial bottle to the adsorbent-filled tube at the pumping/nitrogen flow rates of 50 mL/min. For the human breath samples, we collected the exhaled breath of 10 L from a healthy human using a gas sampling bag (Smart Bag PA CEK-10, GL Science Inc.). Then, the sampling bag was connected to an adsorbentfilled tube, and 500 mL of the collected breath was transferred to the adsorbent-filled tube at the pumping rate of 50 mL/min. For preparing the samples containing chemo-/biomarker molecules, we intentionally introduced the molecule additives including 1-butanol, 2-pentanone, and 1-hexanol for aroma#1, heptanal, 3-octanone, decane, and 3-decanone for aroma#2, 1- pentyn-3-ol, 1-hexanol, heptanal, 3-octanone, and 3-decanone for aroma#3, and heptanal, nonanal, decane, undecane, and benzaldehyde for human breath (as summarized in Table 1). A total of 2 μL of liquid concentrate for each molecule additive was taken in a vial bottle, and the vaporized species was collected together with aroma odor and human breath by using an adsorbent-filled tube. Twenty different samples were prepared for each condition (aroma#1, aroma#2, aroma#3, human breath, and their molecule additive-containing samples). For the human breath samples, we collected the exhaled breath at once from the same donor and divided it into several portions to make sure the reliability of biomarkers without the interference of unexpected biological variations. The sample tubes were sealed and stored in a refrigerator at 4 °C until they were used for GC−MS measurements.

Table 1. Summary of Molecule Additives to Aroma Odor Samples and Human Breath Samples, Which Serve as Chemo-/ Biomarkers in This Study

samples	aroma#1	aroma#2	aroma#3	breath
molecule additives	1-butanol	heptanal	1-pentyn-3-ol	heptanal
	2-pentanone	3-octanone	1-hexanol	nonanal
	1-hexanol	decane	heptanal	decane
		3-decanone	3-octanone	undecane
			3-decanone	benzaldehyde

GC-MS measurement. Mass chromatogram data of the aroma odor samples and the human breath samples were obtained by GC−MS (GCMSQP2020, Shimadzu) using an inlet temperature control unit (OPTIC4). For the aroma odor samples, a SLB-IL60 capillary column (30 m length, 0.25 mm inner diameter, 0.2 μm thickness, Sigma-Aldrich) was used, and the GC oven temperature profile was set as follows: (i) kept constant at 40 °C for 5 min, (ii) increasing to 200 °C at a rate of 10 °C/ min, and (iii) kept at 200 °C for 5 min. For the human breath samples, an InertCap FFAP capillary column (60 m length, 0.25 mm inner diameter, 0.5 µm thickness, GL Science) was used, and the GC oven temperature profile was set as follows: (i) kept constant at 40 °C for 3 min, (ii) increasing to 200 °C at a rate of 5 °C/min, and (iii) kept at 200 °C for 5 min. The inlet temperature was increased to 300 °C with a split flow of He at a rate of 5 mL/min for the aroma odor samples and 2 mL/min for the human breath samples. MS measurements were conducted by electron ionization mode and positive ion analysis. The ion source temperature and the interface temperature at the GC-to-MS junction were 200 °C and 230 °C, respectively. The vacuum pressure was 9.9 × 10−5 Pa. An MS analyzer of single quadrupole and the full scan data acquisition mode were used. Data were analyzed by GCMS Solution ver. 4.45 SP1. The concentrations of chemo-/ biomarker molecules were estimated by calibration curves created using tracer molecules.



	n=1	n=2	...	n=N
d=1	y_{11}	y_{12}	...	y_{1N}
d=2	y_{21}	y_{22}	...	y_{2N}
⋮	⋮	⋮	⋮	⋮
d=D	y_{D1}	y_{D2}	...	y_{DN}

Figure 1. Graphical workflow of NPFimg for visualizing chemo-/biomarker signals from raw gas chromatography−mass spectrometry (GC−MS) data. Starting from (a) series of raw MS data, the workflow follows (b) creation of a primary two-dimensional (2D) MS map (m/z vs retention time) with power-law scale intensity, (c) creation of a regulated 2D MS map by noise reduction and position alignment, (d) image segmentation and machine learning in each segment, (e) feature score calculation, and (f) creation of a 2D feature score map

Data Analysis. Raw MS data were treated and analyzed by the following protocols in NPFimg. The workflow of NPFimg is shown in Figures 1 and S1 and S2. The source codes were developed in Python ver.3.7.7 and are provided on GitHub (http://github.com/poomcj/NPFimg) under the MIT license. First, all MS data, that is, the series of retention time-signal abundance data (.CDF: computable document format) (Figure 1a) were merged and converted into a 2D MS map (.PNG: portable network graphic) as the functions of m/z (xaxis) and retention time (y-axis). The range of m/z and retention time used for analysis were 35−300 (m/z) and 3−26 min (retention time) for aroma odor and 35−300 (m/z) and 3−48 min (retention time) for human breath, respectively. The resolutions of m/z and retention time in raw GC−MS data were 1 and 0.02 s, respectively. The image size of the 2D MS map was 1350 × 3750 pixels, where they correspond to the resolutions of ca. 0.20 in m/z and ca. 0.37 s in retention time for aroma odors and ca. 0.20 in m/z and ca. 0.72 s in retention time for human breath, respectively. For the image processing, the intensity of the 2D MS map (i.e., signal abundance in raw MS data) was scaled by a power law ($\gamma = 0.5$), represented by 256 colors, and normalized via the highest peak using Matplotlib ver.3.2.2 (primary 2D MS map, Figure 1b). A Gaussian filter (SciPy ver1.5.2) based on the dilation method was applied for the noise reduction of the 2D MS map. The position alignment of the 2D MS maps was then performed by identifying the reference peak of external standard−cyclohexene, 1-methyl-4-(1-methyletheny)-(R)- using the blob detection technique[32] and moving window technique,[33] followed by adjusting the reference peak position to be the same in all 2D MS maps (regulated 2D MS map, Figure 1c). After the position alignment, the effective image area of the 2D MS map was 1300 × 3700 pixels. For machine learning, the 2D MS map was divided into the small segments consisting of 1 × 1 or 2 × 2 pixels, and the average intensity of each segment was extracted. Intensity data for the segments with the same address was collected in all 2D MS maps and used as a data set (Figure 1d). In the machine learning process, the data set was divided into training data, validation data, and testing data with the ratio of 50, 25, and 25%, respectively. To enrich the training data set while preventing overfitting, we employed the data augmentation technique. The intensity of the 2D MS maps was randomly modulated in the range of 1.0−10.0% with different interpolation methods including bilinear, hanning, hermite, gaussian, and sinc. Consequently, the number of training data increased by five times the primary ones. The discrimination of the original aroma odor/breath samples and the molecule additive-containing samples and the calculation of the feature score for each data set were performed by the logistic regression model (Figure 1e).[34] Machine learning was performed to optimize the following equation: $\log p/(1 - p) = β0+ x1β1 + x2β2 + x3β3 + ... + xnβn$, where p is the probability of which the data sets can be classified, xn is the intensity of each segment in the 2D MS map, βn ($n \geq 1$) is the model's learned weight (i.e., feature score), and β0 is the bias. The validation data were used to tune the hyperparameters. After obtaining the feature score for all segments in a 2D MS map, a 2D feature score map was created by reconstructing the 2D image with the calculated feature scores at each address (feature score f: $0 \leq f \leq 1$) (Figure 1f). The signals in a 2D feature map were then extracted with their m/z and retention time by blob detection and compared with the MS spectra database (NIST14). In order to confirm the reliability of data analysis in NPFimg, the data analysis was also performed by XCMS and the results were compared. For data analysis by XCMS, peak detection was performed by the CentWave method with the optimized parameter settings.[35] The details of parameter settings are given in Table S1. To evaluate the feature detection performance in XCMS, we counted the number of features by varying the alpha level and optimizing sensitivity and precision. The initially examined alpha level was determined by dividing the highest p-2value obtained in t-test of the detected peaks with the number of examined samples.

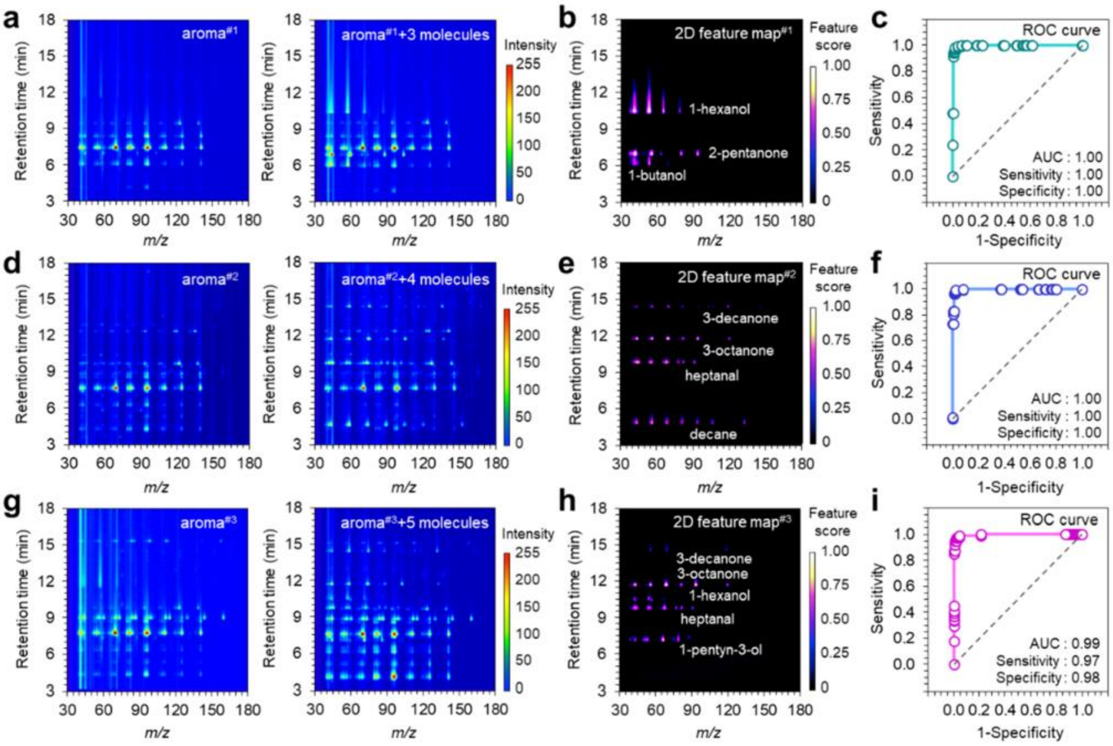


Figure 2. (a,d,g) Regulated 2D MS maps, (b,e,h) 2D feature score maps, and (c,f,i) receiver operating characteristic (ROC) curves of classifiers for (a−c) bergamot organic essential oil—aroma#1 , (d−f) lavender essential oil—aroma#2 , and (g−i) blend essential oil—aroma#3 in comparison with those with chemomarker molecule additives. For the regulated 2D MS maps, the one of original aroma odor is shown in the left and the other with molecule additives is shown in the right. For the visibility, the 2D maps are shown in the restricted range (m/z: 30−180, retention time: 3−18 min). The molecule additives in each sample are summarized in Table 1.

3.4 Results and Discussion

The performance of NPFimg in terms of the identification of multivariate chemomarker features and its time cost is first validated in a case study of aroma odors, which consist of at most 10 species of volatile molecules. Here, we employed three aroma odor samples including bergamot organic essential oil—aroma#1, lavender essential oil—aroma#2, and blend essential oil—aroma#3. We intentionally introduced the molecule additives listed in Table 1 into the original aroma odor samples at the tens parts per million (ppm) order of concentration as the chemomarkers and examined the identification of these molecule additives by comparing them with the original aroma odor samples. Figure 2a shows the 2D MS maps for aroma#1 (i.e., left map) and aroma#1 with three molecule additives (i.e., right map). For the visibility, the 2D MS maps are shown in the restricted range (m/z: 30−180, retention time: 3−18 min). The full range 2D MS maps are shown in Figure S3. The clear difference can be seen in the two maps. Figure 2b shows the 2D feature score map of molecular fragment signals of chemomarkers for discriminating aroma#1 and aroma#1 with molecule additives. For machine learning, the 2D MS map was divided into the segments with the 2 × 2 pixels size because the image quality of the resultant 2D feature score map was comparable to the one with the higher resolution analysis using the segment size of 1 × 1 pixel. Contrary to the 2D MS maps, the 2D feature score map exhibits only the limited number of molecular fragment signals. We confirmed that the addresses of the observed molecular fragment signals on the 2D feature score maps (m/z, retention time) are in good agreement with those of the molecular additives on the 2D MS maps (Figure S4). Figure 2c shows the receiver operating characteristic (ROC) curve of the classifier. The values of area under the curve (AUC), sensitivity and specificity of the classifier are 1.00, 1.00, and 1.00, showing the sufficient reliability of the classifier. Figure 2d−i shows (d,g) the 2D MS maps, (e,h) the 2D feature score maps, and (f,i) the ROC curves for (d,e,f) aroma#2 and aroma#2 with four molecule additives and (g,h,i) aroma#3 and aroma#3 with five molecule additives, respectively. Compared to aroma#1, the larger number of molecular fragment signals is seen in the 2D MS maps of aroma#2, aroma#3, and the ones with molecule additives. We found that the reliable 2D feature score maps (the detailed validation is shown in Figures S5 and S6) and ROC curves were also obtained even when the analytes become more complex (AUC, sensitivity, and specificity were 1.00, 1.00, and 1.00 for aroma#2 and 0.99, 0.97, and 0.98 for aroma#3). With respect to the time cost, the feature identification of chemomarkers in NPFimg was completed within 5 min. The time cost in NPFimg was almost unchanged even when the analytes became complex. Thus, these results clearly validated the performance of NPFimg for the immediate identification of multivariate chemomarker features in analytes.

To evaluate the applicability of NPFimg to more complex analytes, next we examined the ultratrace level biomarker analysis in human breath, which contains over hundreds or thousands of chemical compounds with various concentrations from ppb to ppm orders.[36] We intentionally introduced five molecular additives, including heptanal, nonanal, decane, undecane, and benzaldehyde, into the original exhaled sample at the ppb level as biomarkers. The selected molecules are wellknown lung cancer biomarkers in exhaled breath.[37–40] In order to eliminate the biological variation that influences the reliability of the selected biomarkers, the breath sample was collected from the same donor at once in this study. Figure 3a shows the 2D MS maps for the human breath sample (i.e., left map) and the human breath sample with five molecule additives (i.e., right map).

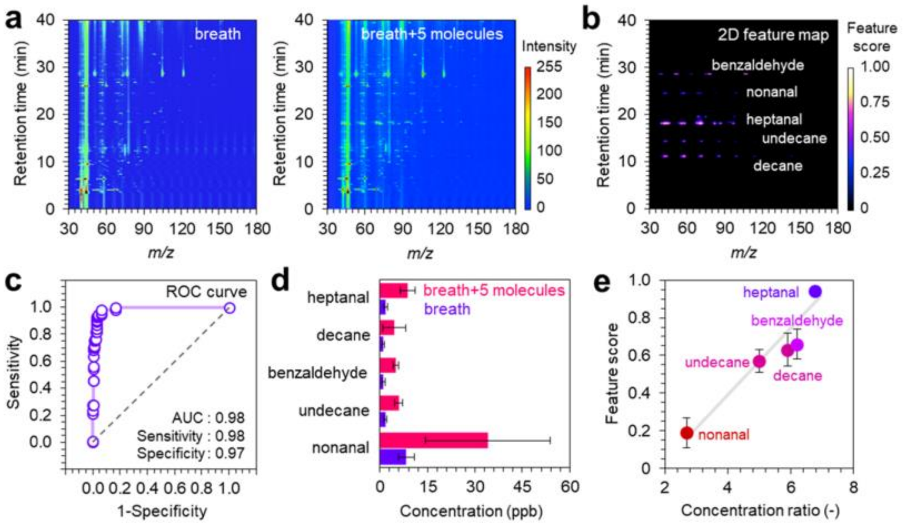


Figure 3. (a) Regulated 2D MS maps, (b) 2D feature score map, and (c) ROC curve of classifiers for breath vs breath + molecule additives, respectively. For the regulated 2D MS maps, the one of original breath is shown in the left and the other with molecule additives is shown in the right. (d) Comparisons in concentrations of biomarker molecules in breath and breath + molecule additives. (e) Relationship between concentration ratio [(breath + molecule additives)/breath] and feature score for the biomarker molecules. For the visibility, the 2D maps are shown in the restricted range (m/z: 30−180, retention time: 3−40 min). The molecule additives are summarized in Table 1.

For the visibility, the 2D maps are shown in the restricted range (m/z: 30−180, retention time: 3−40 min). The full range 2D MS maps are shown in Figure S7. The identification of the differences in the two maps is rather difficult due to their complexities. On the other hand, the 2D feature score map exhibited the limited number of molecular fragment signals, as shown in Figure 3b. We confirmed that the addresses of the molecular fragment signals on the 2D feature score maps are in good agreement with those of the molecular additives on the 2D MS maps (Figure S8). Figure 3c shows the

ROC curve of the classifier. The values of AUC, sensitivity, and specificity of the classifier are 0.98, 0.98, and 0.97, respectively, showing the sufficient reliability of the classifier. The quantitative analysis showed that the concentrations of biomarkers in the analytes were in a few ppb to several tens of ppb level, as shown in Figure 3d. Also, we found that the feature score for each biomarker is critically governed by the concentration ratio in analytes rather than the absolute concentration difference (Figure 3e). This principle allows us to reliably extract the feature of low concentration molecules under the coexistence of high concentration molecules. Thus, these results highlight the applicability of NPFimg to the ultratrace level biomarker analysis in complex analytes, in which both high concentration and low concentration molecules coexist.

We discuss the reliability of MS data processing in NPFimg by comparing it with a widely used analysis software—XCMS. In most of the conventional MS data processing algorithms for characterizing the chemo-/biomarkers, there are two major processes including the peak picking process in the raw MS spectra and the subsequent pairwise peak list comparison process.17−25 We compared NPFimg and XCMS in terms of the performances of signal acquisition in raw MS data and feature identification (Figure S9). The signal acquisition for NPFimg was conducted by the blob detection technique using the regulated MS maps. Totally, 88, 160, and 131 of the molecular fragment peaks were extracted from aroma#1 + three molecule additives, aroma#2 + four molecule additives, and aroma#3 + five molecule additives, respectively, by NPFimg. On the contrary, only 79, 133, and 99 peaks were detected by XCMS. These results are consistent with the previous report, which stated that XCMS produced many false negative peaks (i.e., missing peaks) during its peak picking process.20 The false negative peaks in XCMS would strongly influence the following feature identification process. For the feature identification, contrary to our expectation, a similar number of or more features were identified by XCMS, while it produced false negative peaks during the peak picking process. Sensitivity and precision for feature identification are on average 0.90 and 0.99 in NPFimg and 0.80 and 0.41 in XCMS, respectively, showing the higher ratios of both false positive and false negative features in XCMS. The detailed analysis revealed that the observed false positive and false negative features are caused by the missing peaks during the peak picking process and the batch-to-batch variation of signal intensity in analytes (i.e., batch effect),[41,42] respectively. In order to confirm that these problems are addressed in NPFimg, we evaluated the performance of chemomarker feature identification by varying the functions of NPFimg (Figures S9 and S10). We found that the number of false negative features increased when using the 2D MS map with a linearscale plot, that is, only the limited number of signals can be seen in the map. On the other hand, the number of false positive features increased when removing the intensity normalization process, that is, in case that the signal intensity varies in each batch. These results validated that the problems of missing peaks and the batch effect are successfully addressed by the power-law scale intensity plot and intensity normalization in NPFimg. Nevertheless, the batch effects in untargeted metabolomics/chemometrics need to be carefully corrected by involving other techniques[43] because the intensity normalization used in this study is based on an internal standard, which can be applicable only to quality-controlled biological/chemical replicates. We also found that the false positive features are also produced when the random forest algorithm was used instead of the logistic regression algorithm for machine learning. In this case, false identification of noise as a chemomarker occurred due to its feature identification principle (Figure S11). Thus, the abovementioned results highlight that the functions employed in image processing and the logistic regression algorithm employed in machine learning make NPFimg reliable compared with the conventional peak picking-based data processing approach.

Finally, we demonstrate the applicability of NPFimg to untargeted metabolomics for analyzing the concentration variations of metabolites in analytes (Figure S12). After obtaining the 2D feature score map, the features of biomarkers are extracted by the blob detection technique. The addresses of features are fed back to the regulated 2D MS map with the linear-scale intensity, and the peak area/intensity of the markers is compared among analytes. The MS spectra of the biomarkers (decane, undecane, heptanal, nonanal, and benzaldehyde) showed that the variations of their peak area/ intensity among analytes are successfully observed. Thus, these results demonstrated the feasibility of NPFimg for untargeted metabolomics in complex analytes.

3.5 Conclusions

In conclusion, we presented a method named NPFimg, which automatically identifies multivariate chemo-/biomarkers feature of analytes in chromatography−MS data without the peak picking process, which had been a crucial bottleneck for data processing of raw MS data. NPFimg combines image processing and machine learning and processes a 2D MS map to discriminate analytes and identify and visualize marker features. Our approach allows us to comprehensively characterize the signals in MS data without employing the conventional peak picking process, which suffers from the false peak detections. The feasibility of chemo-/biomarker characterization was successfully demonstrated in case studies of aroma odor and human breath on GC−MS even at the ppb level. Comparison with the widely used XCMS showed the excellent reliability of NPFimg, in that it had lower error rates of the signal acquisition and the feature identification of chemo-/ biomarkers. In addition, we showed the potential applicability of NPFimg to the untargeted metabolomics of human breath. While this study showed the limited applications, NPFimg is potentially applicable to data processing in diverse metabolomics/chemometrics using GC− and LC−MS. Because time cost in NPFimg is much shorter than the peak picking-based conventional approaches, the high throughput online MS data analysis of various complex analytes would be expected by uploading the data file on Cloud space.

3.6 References

1. Eloh, K.; Sasanelli, N.; Maxia, A.; Caboni, P. J. Agric. Food Chem. 2016, 64, 5963−5968.
2. Leite, V. S. A.; Reis, M. R.; Pinto, F. G. ACS Food Sci. Technol. 2021, 1, 242−248.
3. Wang, J.; Jayaprakasha, G. K.; Patil, B. S. ACS Food Sci. Technol. 2021, 1, 77−87.
4. Yao, C. H.; Wang, L.; Stancliffe, E.; Sindelar, M.; Cho, K.; Yin, W.; Wang, Y.; Patti, G. J. Anal. Chem. 2020, 92, 1856−1864.
5. Wikoff, W. R.; Nagle, M. A.; Kouznetsova, V. L.; Tsigelny, I. F.; Nigam, S. K. J. Proteome Res. 2011, 10, 2842−2851.
6. Zhang, W. X.; Li, H. Q.; Xu, Z. D.; Dou, J. J. RSC Adv. 2020, 10, 3092−3104.
7. Meister, I.; Zhang, P.; Sinha, A.; Sköld, C. M.; Wheelock, Å. M.;Izumi, T.; Chaleckis, R.; Wheelock, C. E. Anal. Chem. 2021, 93,5248−5258.
8. Edmands, W. M. B.; Ferrari, P.; Scalbert, A. Anal. Chem. 2014,86, 10925−10931.
9. Alkhalifah, Y.; Phillips, I.; Soltoggio, A.; Darnley, K.; Nailon, W. H.; McLaren, D.; Eddleston, M.; Thomas, C. L. P.; Salman, D. Anal. Chem. 2020, 92, 2937−2945.
10. Bruderer, T.; Gaisl, T.; Gaugg, T. G.; Nowak, N.; Streckenbach, B.; Müller, S.; Moeller, A.; Kohler, M.; Zenobi, R. Chem. Rev. 2019, 119, 10803−10828.
11. Dührkop, K.; Nothias, L. F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, A. M.; Petras, D.; Gerwick, H. W.; Rousu, J.; Dorrestein, C. P.; Bocker, S. Nat. Biotechnol. 2021, 39, 462−471.
12. García-Cañaveras, J. C.; Donato, M. T.; Castell, J. V.; Lahoz, A. J. Proteome Res. 2011, 10, 4825−4834.
13. Thiele, C.; Wunderling, K.; Leyendecker, P. Nat. Methods 2019, 16, 1123−1130.
14. Wang, Z.; Cui, B.; Zhang, F.; Yang, Y.; Shen, X.; Li, Z.; Zhao, W.; Zhang, Y.; Deng, K.; Rong, Z.; Yang, K.; Yu, X.; Li, K.; Han, P.; Zhu, Z. J. Anal. Chem. 2019, 91,

2401−2408.

15. Tsugawa, H.; Nakabayashi, R.; Mori, T.; Yamada, Y.; Takahashi, M.; Rai, A.; Sugiyama, R.; Yamamoto, H.; Nakaya, T.; Yamazaki, M.; Kooke, R.; Bac-Molenaar, J. A.; Oztolan-Erol, N.; Keurentjes, J. J. B.; Arita, M.; Saito, K. Nat. Methods 2019, 16, 295−298.
16. Taylor, M. J.; Lukowski, J. K.; Anderton, C. R. J. Am. Soc. Mass Spectrom. 2021, 32, 872−894.
17. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. Anal. Chem. 2006, 78, 779−787.
18. Forsberg, E. M.; Huan, T.; Rinehart, D.; Benton, H. P.; Warth, B.; Hilmers, B.; Siuzdak, G. Nat. Protoc. 2018, 13, 633−651.
19. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresič, M. ˇBMC Bioinf. 2010, 11, 395.
20. Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. Anal. Chem. 2017, 89, 8689−8695.
21. Tengstrand, E.; Lindberg, J.; Åberg, K. M. Anal. Chem. 2014, 86, 3435−3442.
22. Ji, H.; Zeng, F.; Xu, Y.; Lu, H.; Zhang, Z. Anal. Chem. 2017, 89, 7631−7640.
23. Wanichthanarak, K.; Fan, S.; Grapov, D.; Barupal, D. K.; Fiehn, O. PLoS One 2017, 12, No. e0171046.
24. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; Vandergheynst, J.; Fiehn, O.; Arita, M. Nat. Methods 2015, 12, 523−526.
25. O'Shea, K.; Misra, B. B. Metabolomics 2020, 16, 36.
26. Borgsmüller, N.; Gloaguen, Y.; Opialla, T.; Blanc, E.; Sicard, E.; Royer, A.-L.; Bizec, B. L.; Durand, S.; Mingné, C.; Pétéra, M.; PujosGuillot, E.; Giacomoni, F.; Guitton, Y.; Beule, D.; Kirwan, J. Metabolites 2019, 9, 171.
27. Woldegebriel, M.; Vivó-Truyols, G. Anal. Chem. 2015, 87, 7345−7355.
28. Wiczling, P.; Kamedulska, A.; Kubik, Ł. Anal. Chem. 2021, 93, 6961−6971.
29. Liu, Z.; Portero, E. P.; Jian, Y.; Zhao, Y.; Onjiko, R. M.; Zeng, C.; Nemes, P. Anal.

Chem. 2019, 91, 5768−5776.

30. Melnikov, A. D.; Tsentalovich, Y. P.; Yanshole, V. V. Anal. Chem. 2020, 92, 588−592.

31. Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. Metabolites 2020, 10, 243.

32. Marsh, B. P.; Chada, N.; Sanganna Gari, R. R.; Sigdel, K. P.; King, G. M. Sci. Rep. 2018, 8, 978.

33. Fu, J.; Chu, W.; Dixson, R.; Orji, G.; Vorburger, T. AIP Conf. Proc. 2009, 1173, 280−284.

34. Chai, H.; Liang, Y.; Wang, S.; Shen, H.-W. Sci. Rep. 2018, 8, 13009.

35. Manier, S. K.; Keller, A.; Meyer, M. R. Drug Test. Anal. 2019, 11, 752−761.

36. Chan, L. W.; Anahtar, M. N.; Ong, T. H.; Hern, K. E.; Kunz, R. R.; Bhatia, S. N. Nat. Nanotechnol. 2020, 15, 792−800.

37. Fuchs, P.; Loeseken, C.; Schubert, J. K.; Miekisch, W. Int. J. Cancer 2010, 126, 2663−2670.

38. Chen, X.; Xu, F.; Wang, Y.; Pan, Y.; Lu, D.; Wang, P.; Ying, K.; Chen, E.; Zhang, W. Cancer 2007, 110, 835−844.

39. Bouza, M.; Gonzalez-Soto, J.; Pereiro, R.; De Vicente, J. C.; Sanz-Medel, A. J. Breath Res. 2017, 11, No. 016015.

40. Peng, G.; Hakim, M.; Broza, Y. Y.; Billan, S.; Abdah-Bortnyak, R.; Kuten, A.; Tisch, U.; Haick, H. Br. J. Cancer 2010, 103, 542−551.

41. Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z. J.; Li, Z.; Li, K. Anal. Chem. 2020, 92, 5082−5090.

42. Liu, Q.; Walker, D.; Uppal, K.; Liu, Z.; Ma, C.; Tran, V.; Li, S.; Jones, D. P.; Yu, T. Sci. Rep. 2020, 10, 13856.

43. Wehrens, R.; Hageman, J. A.; van Euwijk, F.; Kooke, R.; Flood, P. J.; Wijnker, E.; Keurentjes, J. J. B.; Lommen, A.; van Eekelen, H. D. L. M.; Hall, R. D.; Munn, R.;

de Vos, R. C. H. Metabolomics 2016, 12, 88.

3.7 Supporting Information

Workflow Details of *NPFimg*

NPFimg is developed for automatically identifying the chemo/bio-makers features of analytes in chromatography-MS data without conventional peak-picking process, which had been a crucial bottleneck in MS data analysis. Firstly, all MS data are merged and converted into a 2D MS map (*m/z* vs. retention time). Next, image processing, which includes i) power-law scale plot, ii) intensity normalization, iii) noise reduction and iv) position alignment, is performed to regulate the 2D MS map to be used in machine learning. A primary 2D MS map is obtained by performing power-law scale plot and intensity normalization. The power-law scale plot allows us to enhance the visibility of low-intensity signals and process all signals with a wide range of intensities together on the map. For the intensity normalization, the intensity of all signals is normalized by that of an internal standard with the highest intensity. A regulated 2D MS map is obtained by performing noise reduction and position alignment. Above-mentioned power-law scale plot enhances not only the visibility of low-intensity signals from analyte but also that of noise from the measurement system. Since high visibility noise sometimes leads to the overlearning in machine learning process, it should be removed here. The spike noise and the streak noise can be successfully eliminated via the Gaussian filtering with maintaining the low-intensity signals (Figure S1). Subsequently, the position alignment is performed to correct the shifts of data points in each 2D MS map. In GC- and LC-MS, the shifts of data points mainly occur along the direction of retention time, which is frequently caused by the degradation of stationary phases in GC/LC columns.[1] In this study, such shifts are corrected by blob detection technique[2] and moving window technique[3] using an external standard. The standard deviations of the peak positions can be successfully reduced to ca. 2×10^{-3} min after the position alignment (Figure S2). For machine learning, the regulated 2D MS map is further divided into small segments consisting of 1×1 or 2×2 pixels, and the intensity of segments at the same address are used as the datasets for training, validating and testing a model. The logistic regression algorithm[4] is used to classify the datasets and calculate a feature score, and finally a 2D feature score map is obtained by reconstructing the 2D image with the calculated feature scores at each address. Since the pixel size employed for the calculation is much smaller than the width of MS peak, the signal acquisition performance of our approach can be higher than that of the peak-picking based data processing. Also, the analysis in *NPFimg* can be completed quickly because of the small data size of 2D image (~200 kB) in *NPFimg* than that of raw MS data (~40 MB). Thus, *NPFimg* may identify the multivariate features of potential chemo/bio-markers immediately without peak-picking process, which was inevitable data processing task in previous studies.

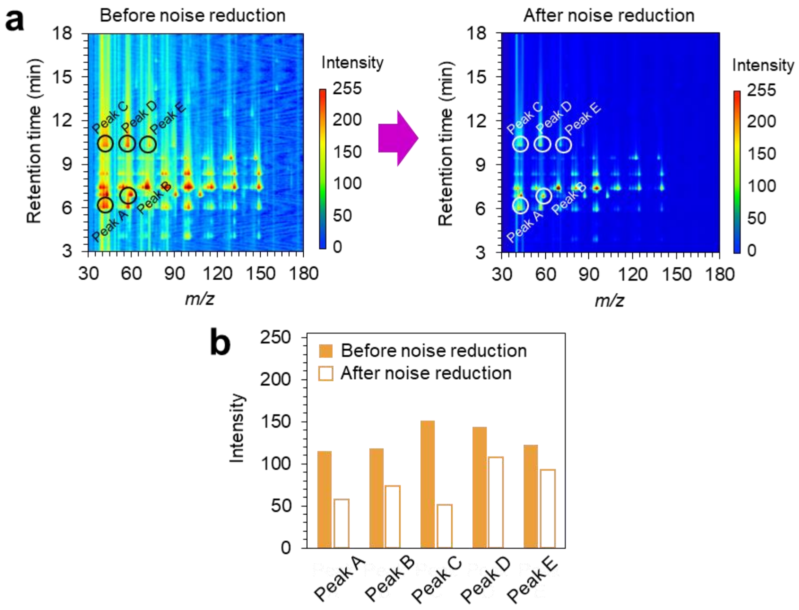
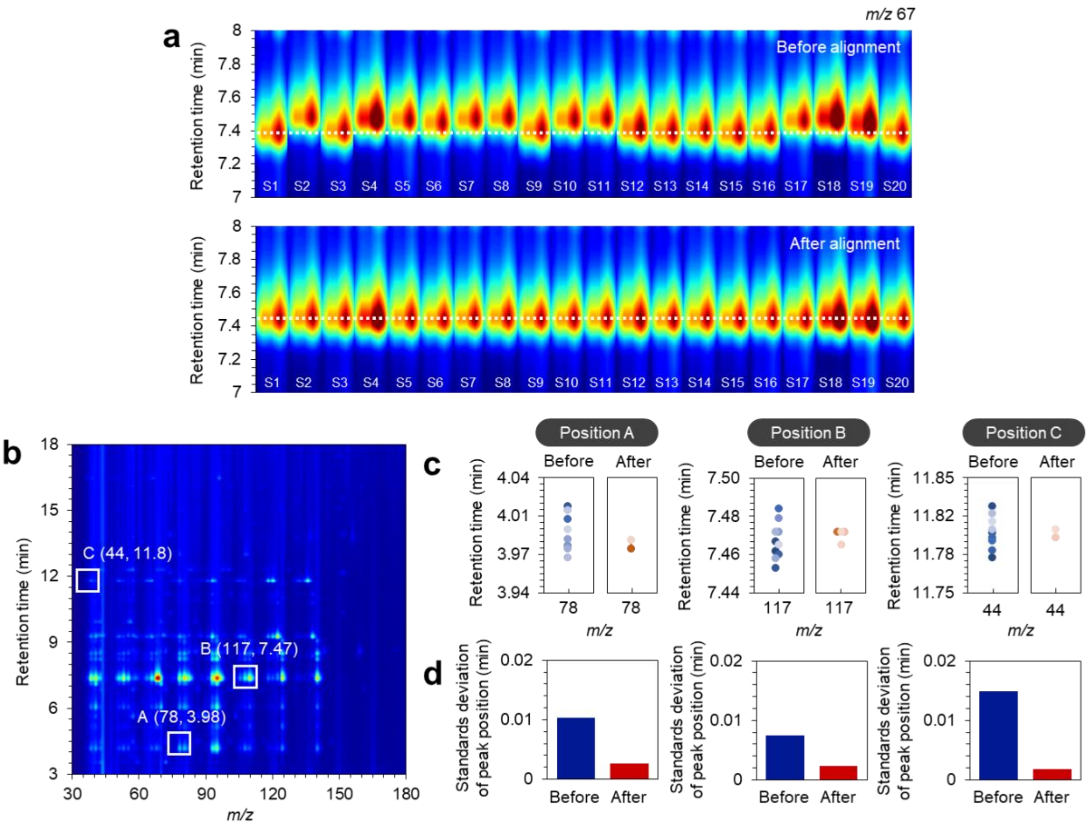


Figure S1. (a) Primary 2D MS maps of aroma#1+three molecule additives with power-law scaled intensity before (left) and after (right) noise reduction via Gaussian filtering. (b) Intensity comparison of the five selected molecular fragment peaks before and after noise reduction. The positions of selected molecular fragment peaks are indicated in (a). The intensity of molecular fragment signals averagely decreases by ~40.8 ± 15.8%.



e

Position	S.D before alignment (min)	S.D. after alignment (min)	Decreasing ratio of S.D. (%)
1	0.0112	0.0034	69.55
2	0.0110	0.0026	76.43
3	0.0095	0.0022	76.95
4	0.0113	0.0033	70.71
5	0.0103	0.0026	74.65
6	0.0102	0.0000	100.00
7	0.0069	0.0031	55.39
8	0.0075	0.0031	58.85
9	0.0075	0.0024	68.37
10	0.0081	0.0024	70.69
11	0.0081	0.0017	78.53
12	0.0148	0.0033	77.73
13	0.0147	0.0033	77.66
14	0.0147	0.0024	83.43
15	0.0142	0.0000	100.00
16	0.0149	0.0018	87.97
17	0.0150	0.0034	77.52
18	0.0142	0.0031	78.01
19	0.0142	0.0000	100.00
20	0.0115	0.0023	79.81
Average			78.11

Figure S2. (a) The extracted reference peak of cyclohexene, 1-methyl-4-(1-methyletheny)-(R)- in 2D MS maps of 20 samples before (upper) and after the position alignment (lower). (b-c) The results of the position alignment in 2D MS maps. (b) The selected peaks in 2D MS map, (c) the positions of the selected peaks in 20 samples before (left) and after (right) the position alignment. (d) The standard deviation values of peak position of the selected peaks before and after the peak alignment. (e) The results of the position alignment for the randomly selected 20 peaks in the 2D MS map. S.D. stands for the standard deviation value.

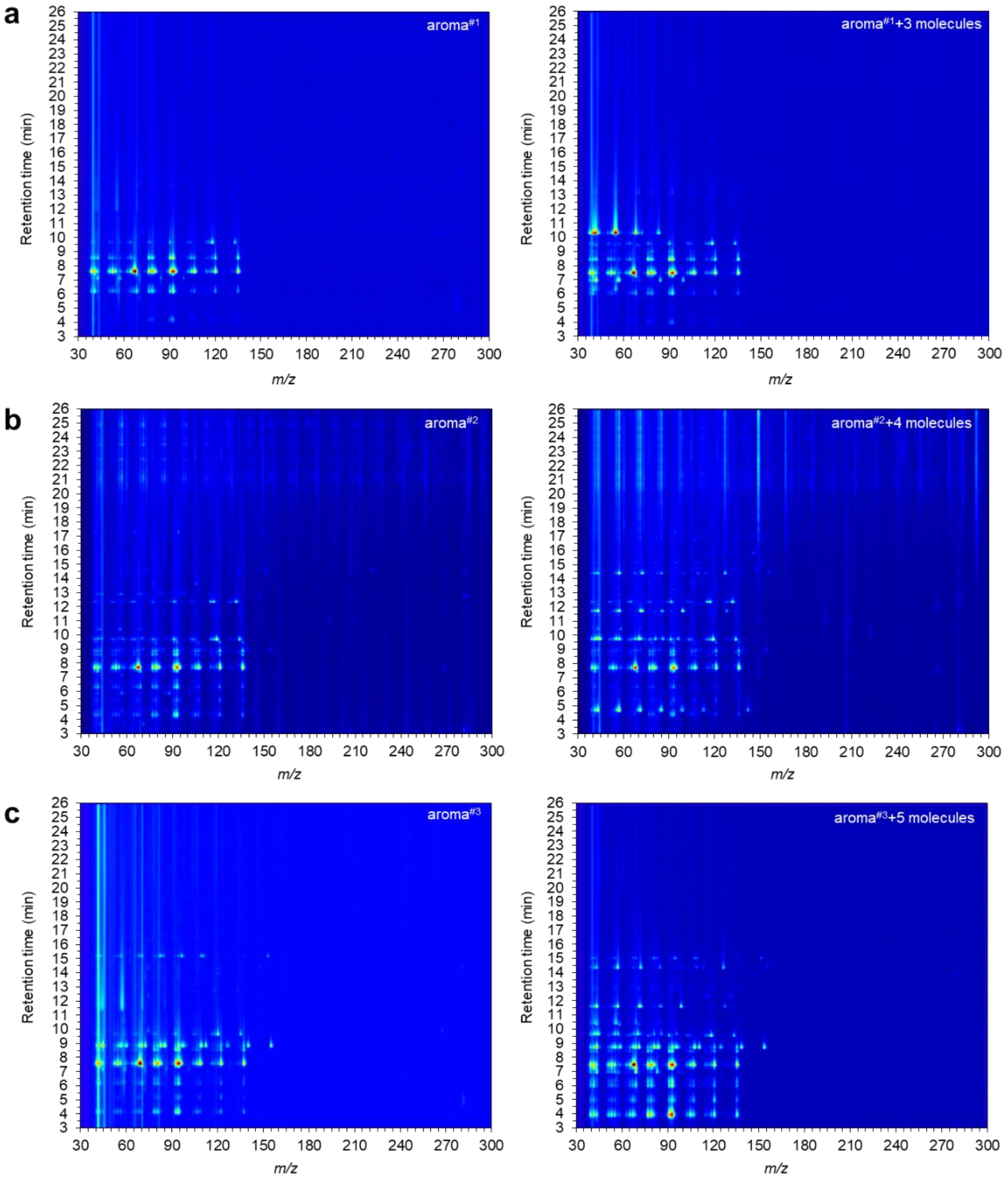


Figure S3. The full range regulated 2D MS maps of (a) aroma#1 (left) and aroma#1 + three molecule additives (right), (b) aroma#2 (left) and aroma#2 + four molecule additives (right) and (c) aroma#3 (left) and aroma#3 + five molecule additives (right), respectively. The molecule additives are 1-butanol, 2-pentanone, 1-hexanol for aroma#1, heptanal, 3-octanone, decane, 3-decanone for aroma#2, and 1-pentyn-3-ol, 1-hexanol, heptanal, 3-octanone, 3-decanone for aroma#3, respectively.

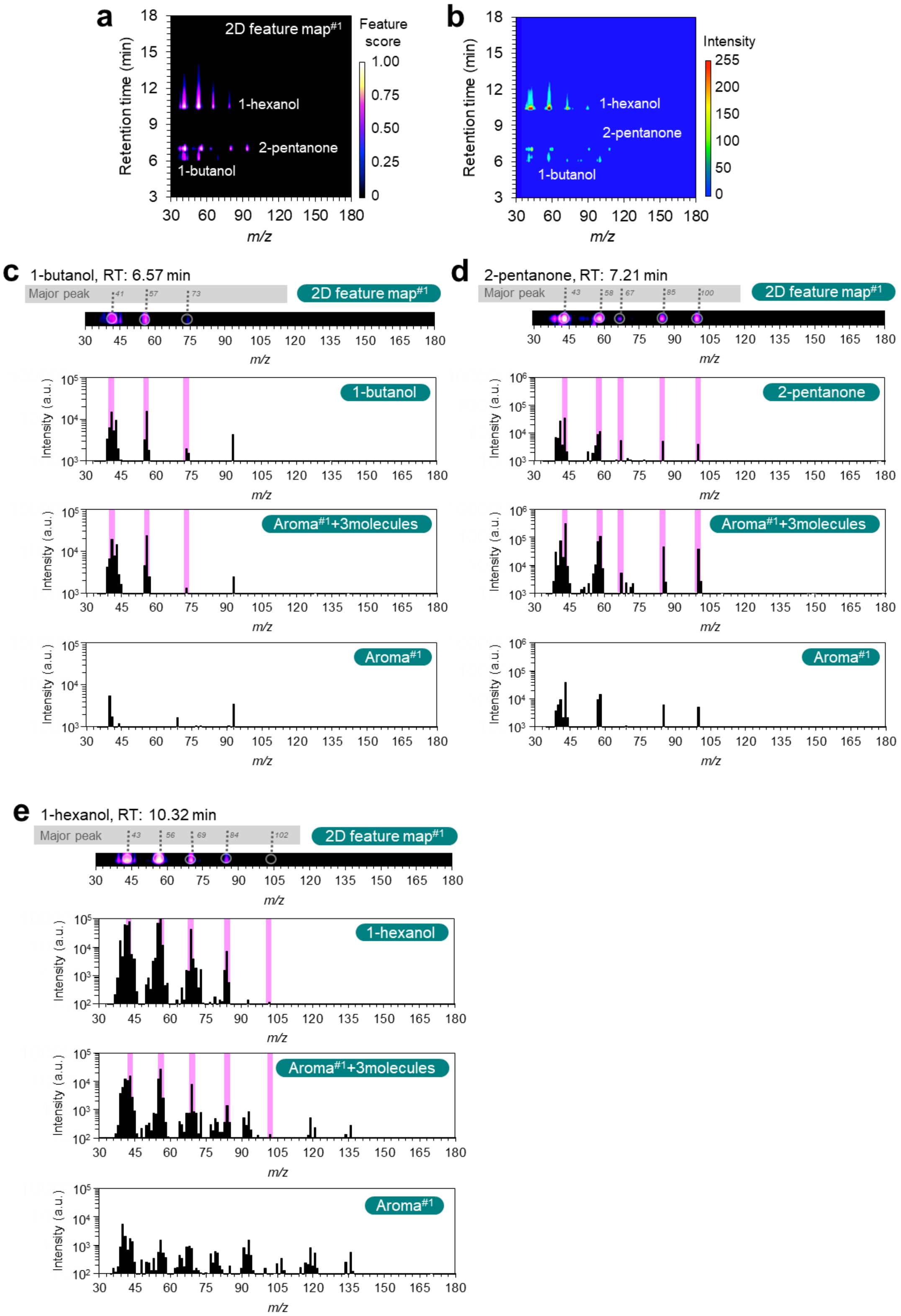


Figure S4. (a) 2D feature score map of aroma#1 vs. aroma#1 + three molecule additives and (b) 2D MS map of chemomarkers (1-butanol, 2-pentanone, 1-hexanol).

(c-e) Comparison of molecular fragment peaks in 2D feature score map and MS data for (c) 1-butanol, (d) 2-pentanone, and (e) 1-hexanol. Upper MS data: the specific chemomarker molecule, middle MS data: aroma + additive molecules, and lower MS data: aroma, respectively.

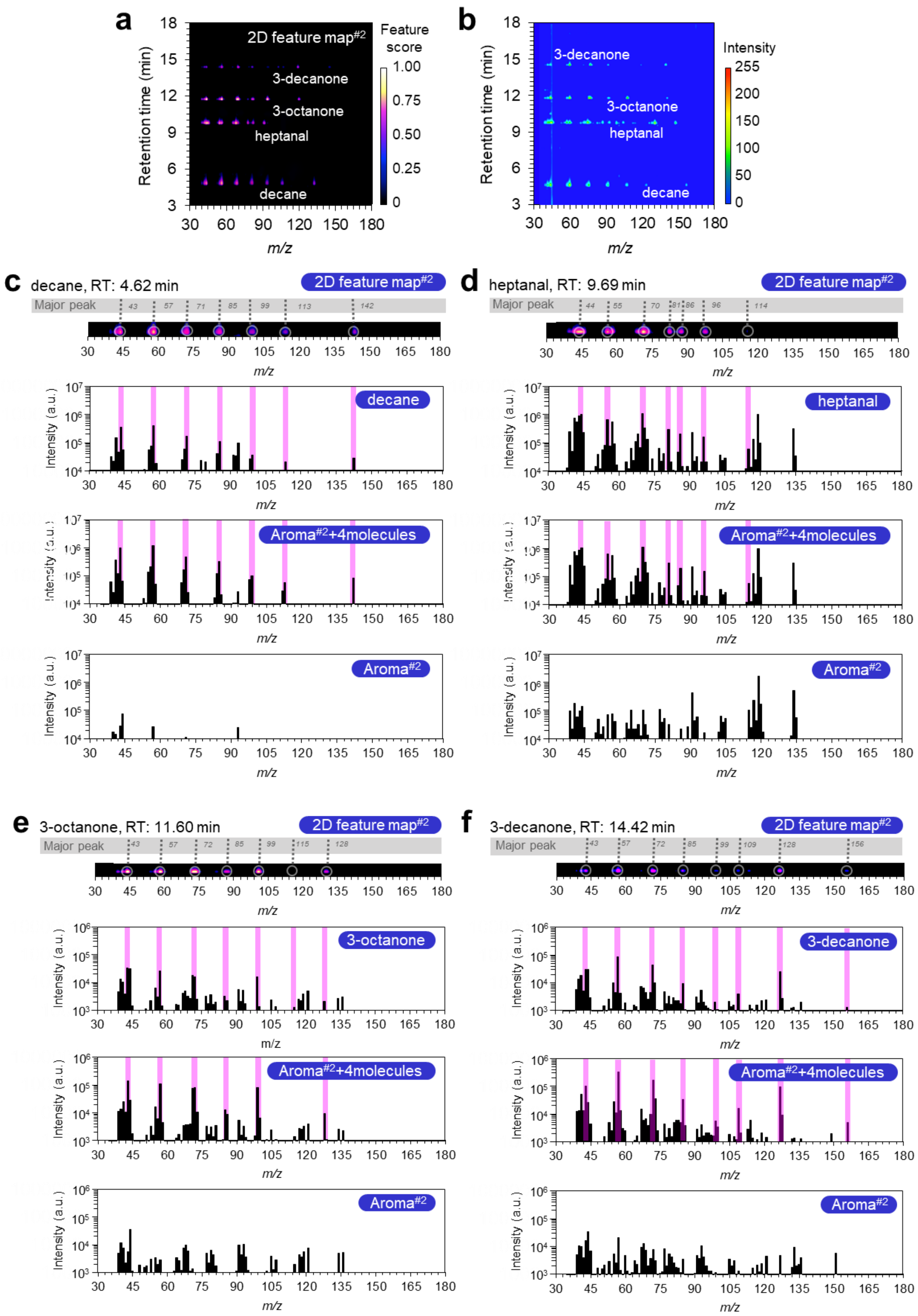
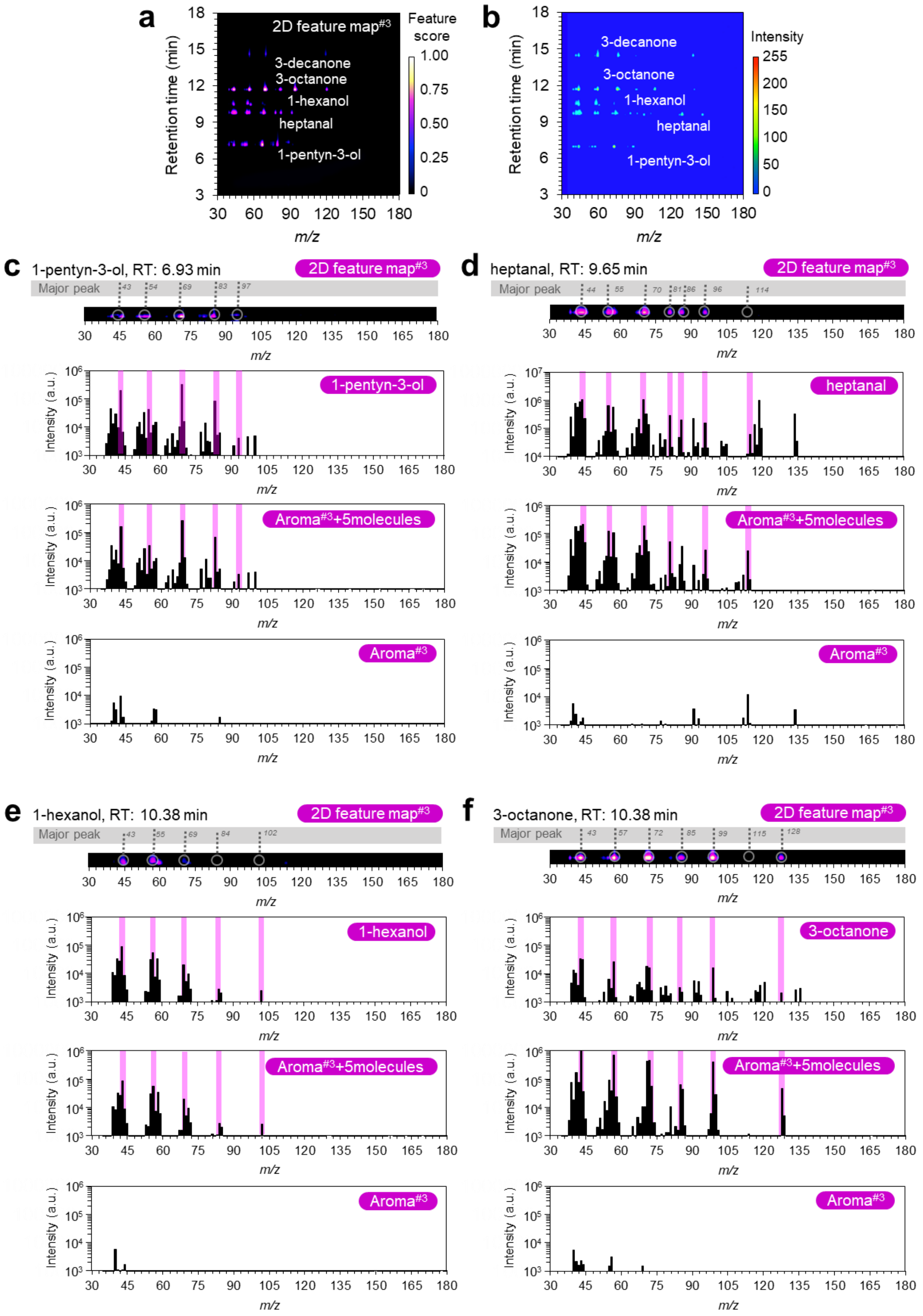


Figure S5. (a) 2D feature score map of aroma#2 vs. aroma#2 + four molecule additives and (b) 2D MS map of chemomarkers (heptanal, 3-octanone, decane, 3-

decanone). (c-f) Comparison of molecular fragment peaks in 2D feature score map and MS data for (c) heptanal, (d) 3-octanone, (e) decane and (f) 3-decanone. Upper MS data: the specific chemomarker molecule, middle MS data: aroma + additive molecules, and lower MS data: aroma, respectively.


a
2D feature map#3
Feature score
1.00
0.75
0.50
0.25
0
3-decanone
3-octanone
1-hexanol
heptanal
1-pentyn-3-ol
Retention time (min)
18
15
12
9
6
3
30 60 90 120 150 180
m/z
b
Intensity
255
200
150
100
50
0
c
1-pentyn-3-ol, RT: 6.93 min
Major peak
43
54
69
83
97
d
heptanal, RT: 9.65 min
44
55
70
81
86
96
114
e
1-hexanol, RT: 10.38 min
43
55
69
84
102
f
3-octanone, RT: 10.38 min
43
57
72
85
99
115
128
Aroma#3+5molecules
Aroma#3
Intensity (a.u.)
10^3
10^4
10^5
10^6
10^7
30 45 60 75 90 105 120 135 150 165 180

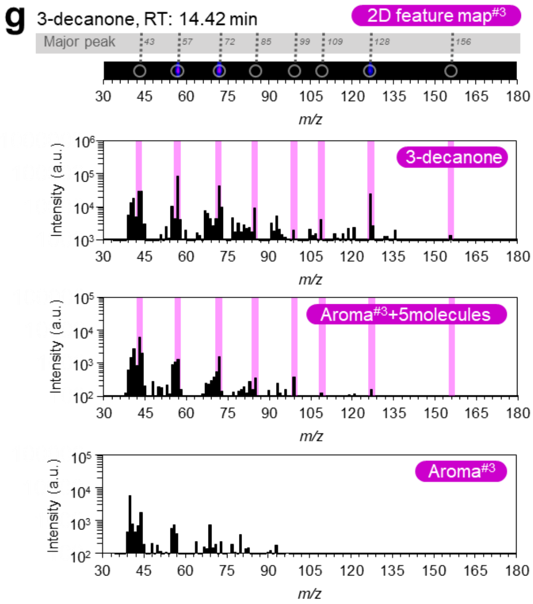


Figure S6. (a) 2D feature score map of aroma#3 vs. aroma#3 + five molecule additives and (b) 2D MS map of chemomarkers (1-pentyn-3-ol, 1-hexanol, heptanal, 3-octanone, 3-decanone). (c-g) Comparison of molecular fragment peaks in 2D feature score map and MS data for (c) 1-pentyn-3-ol, (d) 1-hexanol, (e) heptanal, (f) 3-octanone and (g) 3-decanone. Upper MS data: the specific chemomarker molecule, middle MS data: aroma + additive molecules, and lower MS data: aroma, respectively.

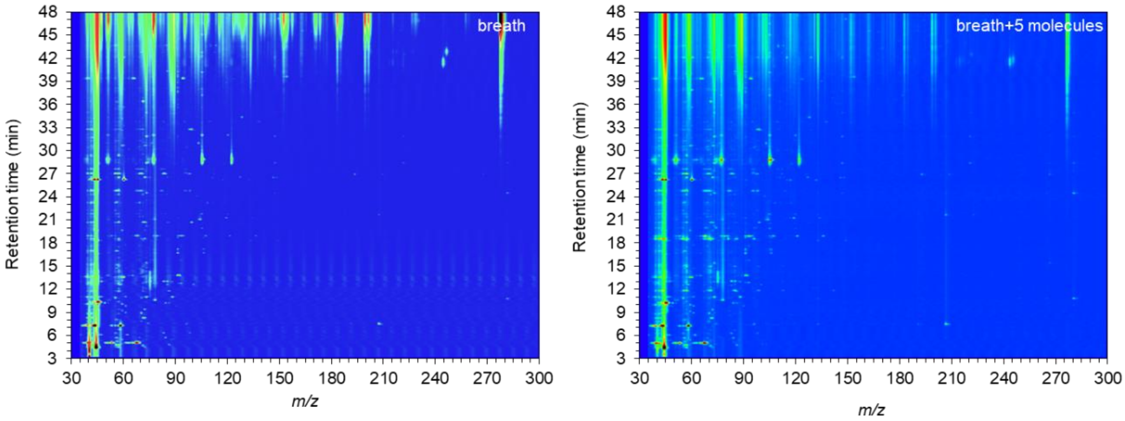
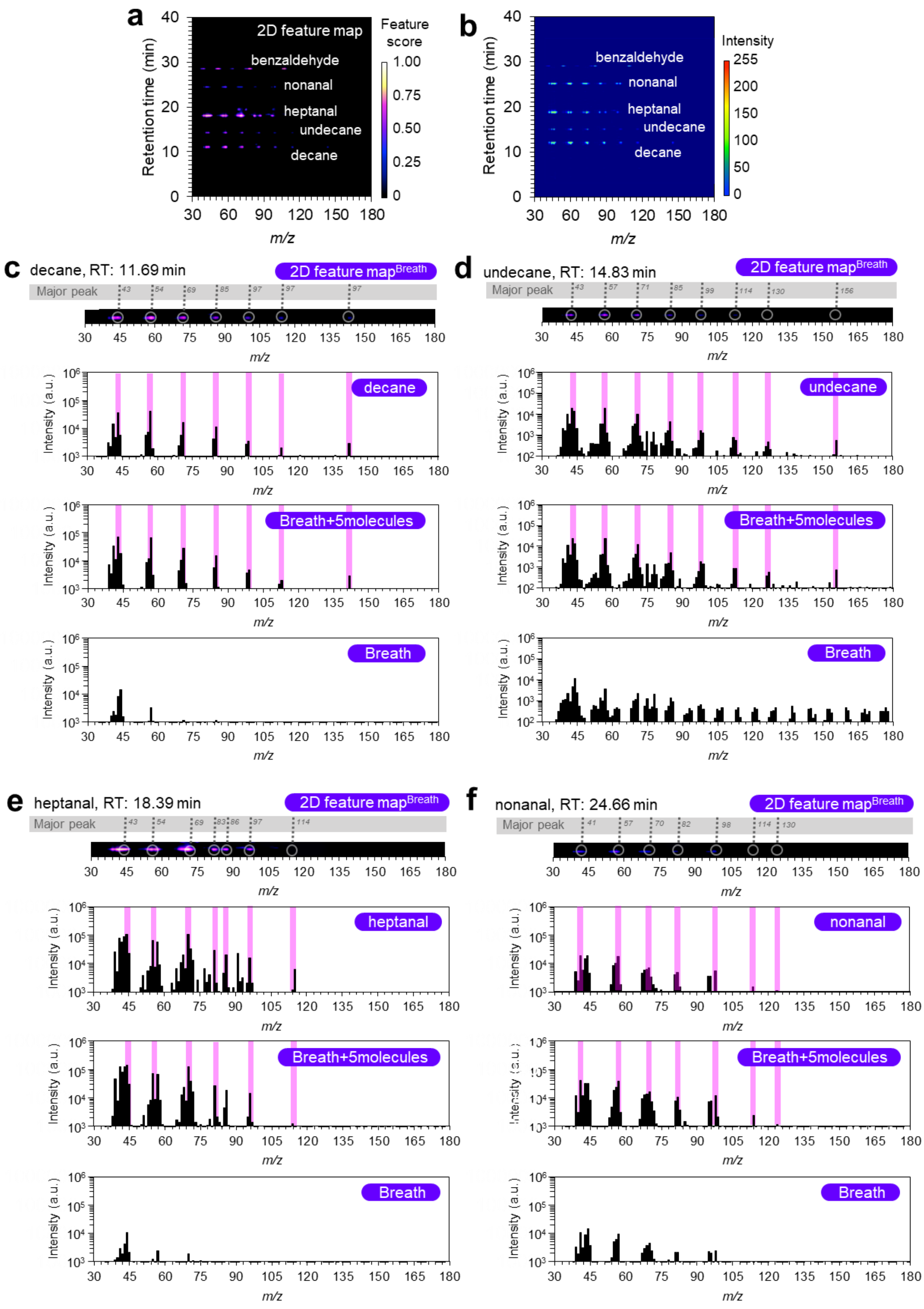


Figure S7. The full range regulated 2D MS maps of breath (left) and breath + molecule additives (right). The molecule additives are heptanal, nonanal, decane, undecane, benzaldehyde.


a
2D feature map
Retention time (min)
m/z
Feature score
1.00
0.75
0.50
0.25
0
benzaldehyde
nonanal
heptanal
undecane
decane
b
Intensity
255
200
150
100
50
0
benzaldehyde
nonanal
heptanal
undecane
decane
c
decane, RT: 11.69 min
2D feature map^{Breath}
Major peak
43
54
69
85
97
97
97
decane
Breath+5molecules
Breath
Intensity (a.u.)
d
undecane, RT: 14.83 min
2D feature map^{Breath}
Major peak
43
57
71
85
99
114
130
156
undecane
Breath+5molecules
Breath
e
heptanal, RT: 18.39 min
2D feature map^{Breath}
Major peak
43
54
69
83
86
97
114
heptanal
Breath+5molecules
Breath
f
nonanal, RT: 24.66 min
2D feature map^{Breath}
Major peak
41
57
70
82
98
114
130
nonanal
Breath+5molecules
Breath

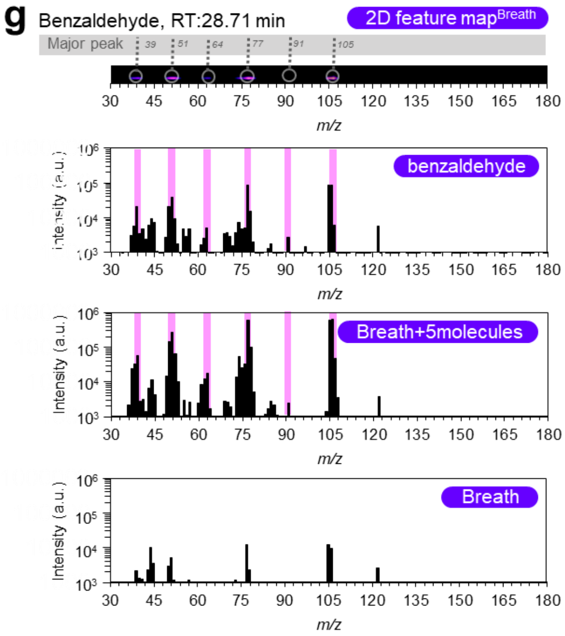
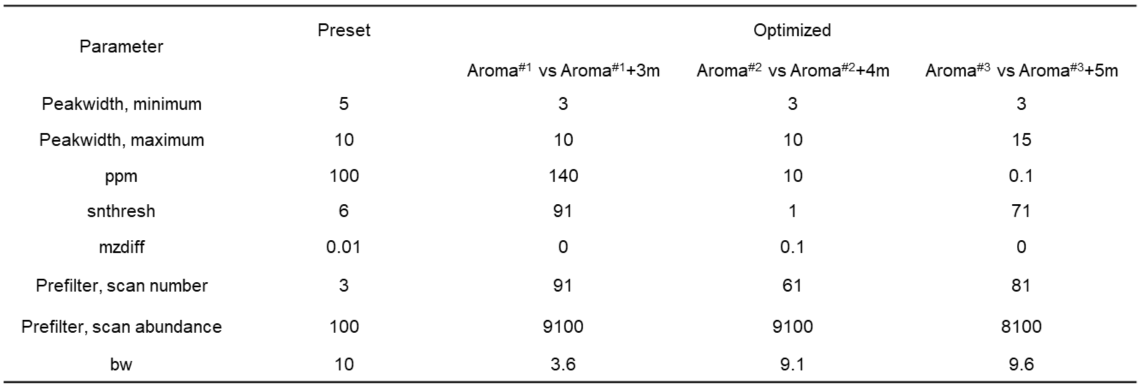


Figure S8. (a) 2D feature score map of breath vs. breath + five molecule additives and (b) 2D MS map of biomarkers (heptanal, nonanal, decane, undecane, benzaldehyde). (c-g) Comparison of molecular fragment peaks in 2D feature score map and MS data for (c) heptanal, (d) nonanal, (e) decane, (f) undecane and (g) benzaldehyde. Upper MS data: the specific chemomarker molecule, middle MS data: aroma + additive molecules, and lower MS data: aroma, respectively.

Table S1. The preset *XCMS* parameters and the optimized *XCMS* parameters.



Parameter	Preset	Optimized		
		Aroma$^{\#1}$ vs Aroma$^{\#1}$+3m	Aroma$^{\#2}$ vs Aroma$^{\#2}$+4m	Aroma$^{\#3}$ vs Aroma$^{\#3}$+5m
Peakwidth, minimum	5	3	3	3
Peakwidth, maximum	10	10	10	15
ppm	100	140	10	0.1
snthresh	6	91	1	71
mzdiff	0.01	0	0.1	0
Prefilter, scan number	3	91	61	81
Prefilter, scan abundance	100	9100	9100	8100
bw	10	3.6	9.1	9.6

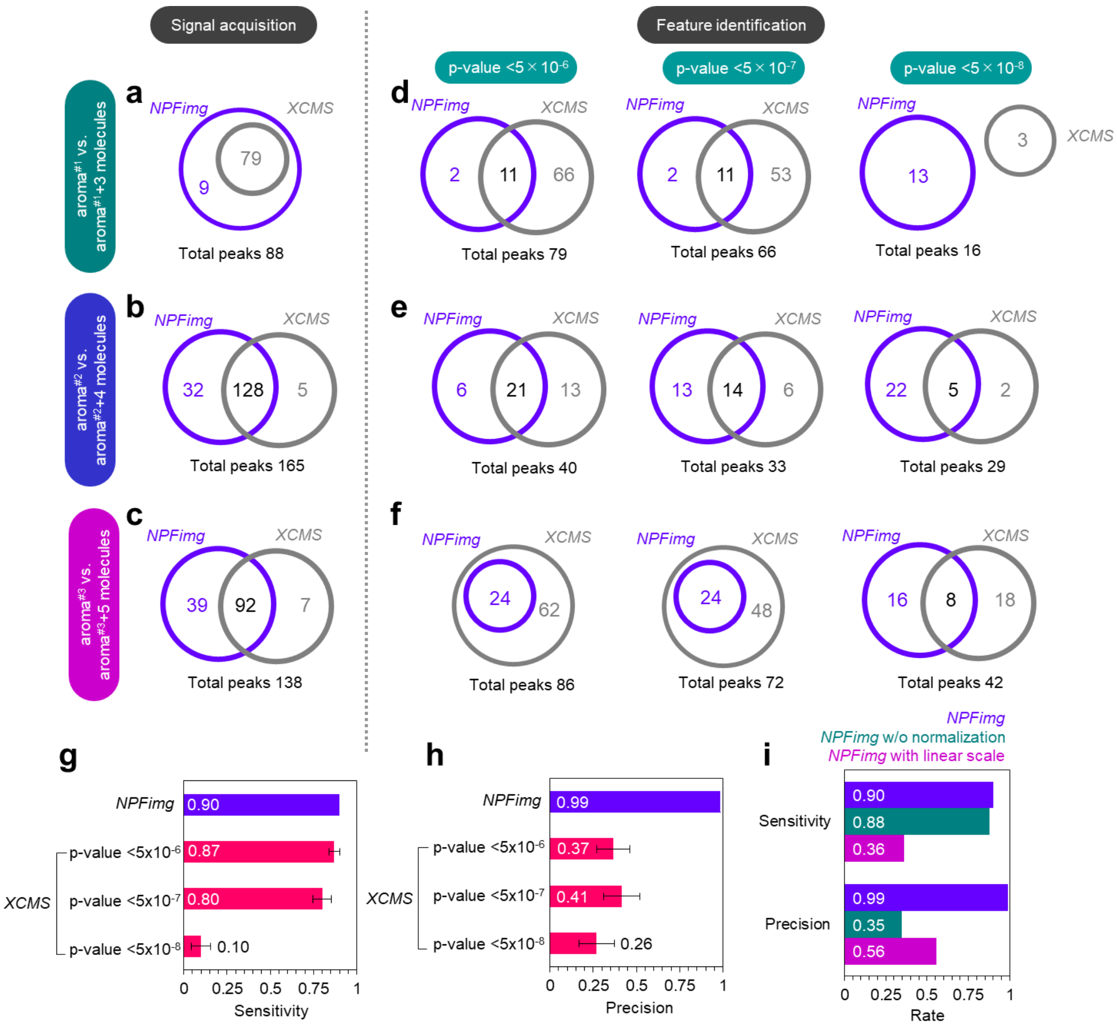


Figure S9. (a-c) Venn diagrams for the number of detected molecular fragment peaks using *NPFimg* and *XCMS* for (a) aroma#1 + three molecule additives, (b) aroma#2 + four molecule additives and (c) aroma#3 + five molecule additives, respectively. (d-f) Venn diagrams for the number of identified chemomarker features using *NPFimg* and *XCMS* for (d) aroma#1 vs. aroma#1 + three molecule additives, (e) aroma#2 vs. aroma#2 + four molecule additives and (f) aroma#3 vs. aroma#3 + five molecule additives, respectively. The molecule additives for each aroma sample are summarized in Table 1. (g,h) Comparison of *NPFimg* and *XCMS* in terms of the performance of chemomarkers feature identification in aroma vs. aroma + molecule additives. (i) Comparison of *NPFimg* with various conditions in terms of the performance of chemomarkers feature identification in aroma vs. aroma + molecule additives (*NPFimg*, *NPFimg* without intensity normalization and *NPFimg* with linear scale intensity plot). The values of sensitivity and precision shown in (g-i) are the averaged ones for the results of three analyses (aroma#1, aroma#2, aroma#3). The number of real features was obtained by comparing the MS data of chemomarkers and database (see Figure S4-S6).

We discuss the reliability of MS data processing in *NPFimg* by comparing with a widely used analysis software—*XCMS*. We compared *NPFimg* and *XCMS* in terms of the performances of signal acquisition in raw MS data and feature identification. The signal acquisition for *NPFimg* was conducted by the blob detection technique using the regulated MS maps. For the data analysis by *XCMS*, the peak detection was performed by *CentWave* method with the optimized parameter settings of *ppm, peakwidth minimum, peakwidth maximum, snthresh, mzdiff, prefilter scan number, prefilter scan abundance* and *bw* (see Table S1). Figure S9a-c show the Venn diagrams for the number of extracted peaks in the raw MS map/spectra using *NPFimg* and *XCMS* for the aroma odor analytes ((a) aroma#1 + three molecule additives, (b) aroma#2 + four molecule additives, (c) aroma#3 + five molecule additives). Totally, 88, 160 and 131 of the molecular fragment peaks were extracted from aroma#1 + three molecule additives, aroma#2 + four molecule additives, and aroma#3 + five molecule additives, respectively, by *NPFimg*. On the contrary, only 79, 133 and 99 peaks were detected by *XCMS*. These results are consistent with the previous report that *XCMS* produced many false negative peaks (*i.e.* missing peaks) during its peak-picking process.5 The false negative peaks in *XCMS* would strongly influence the following feature identification process. Figure S9d-f show the Venn diagrams for the number of identified chemomarker features for the aroma odor samples using *NPFimg* and *XCMS* ((d) aroma#1 vs. aroma#1 + three molecule additives, (e) aroma#2 vs. aroma#2 + four molecule additives, (f) aroma#3 vs. aroma#3 + five molecule additives). On the contrary to our expectation, the similar or larger number of features were identified by *XCMS* especially at higher alpha level while it produced false negative peaks during the peak-picking process (Figure S9a-c). Figure S9g-h shows the comparison of *NPFimg* and *XCMS* in terms of the reliablity of chemomarker features identification in aroma vs. aroma + molecule additives, which were evaluated by comparing the identified features and the raw MS data of chemomarkers. Sensitivity and precision are averagely 0.90 and 0.99 in *NPFimg*, and 0.80 and 0.41 in *XCMS* (at alpha level of 5×10-7), respectively, showing the higher ratios of both false positive and false negative features in *XCMS*. The detailed analysis revealed that the observed false positive and false negative features are caused by the missing peaks during peak-picking process and the batch-to-batch variation of signal intensity in analytes (*i.e.* batch effect), respectively. In order to confirm that these problems are addressed in *NPFimg*, we evaluated the performance of chemomarker features identification by varying the functions of *NPFimg* in Figure S9i and Figure S10. We found that the number of false negative feature increased when using the 2D MS map with a linear-scale plot, *i.e.* only the limited number of signals can be seen in the map. On the other hand, the number of false positive feature increased when removing the intensity normalization process, *i.e.* in case that the signal intensity varies in each batch. These results validated that the problems of missing peaks and the batch effect are successfully addressed by the power-law scale intensity plot and the intensity normalization in *NPFimg*. We also found that the false positive features are also produced when random forest algorithm was used instead to logistic regression algorithm for machine learning. In this case, the false identification of noise as a chemomarker occurred due to its feature identification principle (Figure S11). Thus, above results highlight that the functions employed in image processing and the logistic regression algorithm employed in machine learning make *NPFimg* reliable compared with the conventional peak-picking based data processing approach.

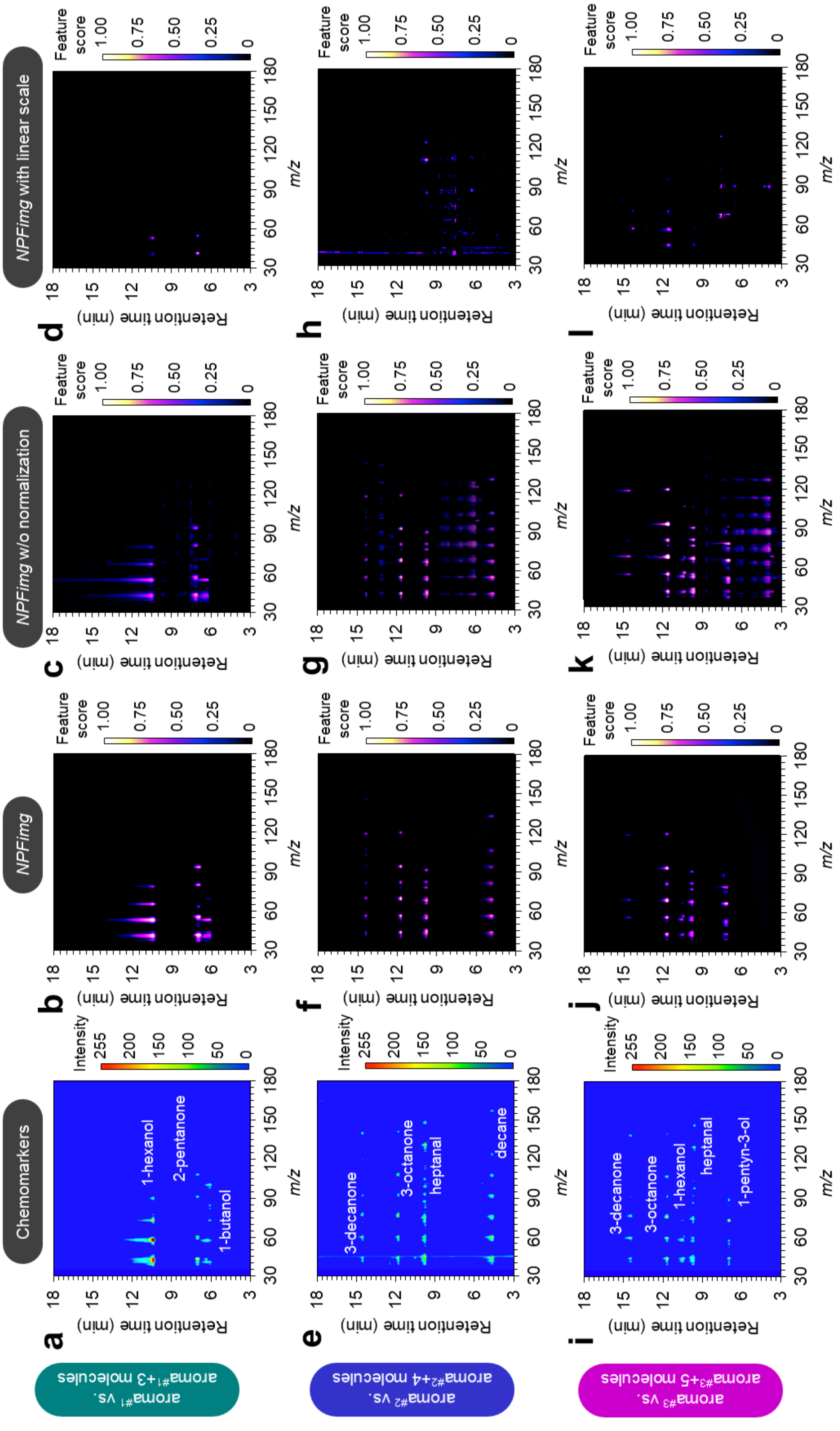

Chemomarkers
NPFimg
NPFimg w/o normalization
NPFimg with linear scale
aroma#1 vs. aroma#1+3 molecules
aroma#2 vs. aroma#2+4 molecules
aroma#3 vs. aroma#3+5 molecules
a
b
c
d
e
f
g
h
i
j
k
l
Retention time (min)
m/z
18 15 12 9 6 3
30 60 90 120 150 180
Intensity
255 200 150 100 50 0
Feature score
1.00 0.75 0.50 0.25 0
1-hexanol
2-pentanone
1-butanol
3-decanone
3-octanone
heptanal
decane
3-decanone
3-octanone
1-hexanol
heptanal
1-pentyn-3-ol

Figure S10. (a) 2D MS map of chemomarkers and (b-d) 2D feature score maps of aroma#1 vs. aroma#1 + three molecule additives, obtained by (b) *NPFimg*, (c) *NPFimg* without intensity normalization process, and (d) *NPFimg* with linear scale plot, respectively. (e) 2D MS map of chemomarkers and (f-h) 2D feature score maps of aroma#2 vs. aroma#2 + four molecule additives, obtained by (f) *NPFimg*, (g) *NPFimg* without intensity normalization process, and (h) *NPFimg* with linear scale plot, respectively. (i) 2D MS map of chemomarkers and (j-l) 2D feature score maps of aroma#3 vs. aroma#3 + five molecule additives, obtained by (j) *NPFimg*, (k) *NPFimg* without intensity normalization process, and (l) *NPFimg* with linear scale plot, respectively. For the visibility, the 2D maps are shown in the restricted range (*m/z*: 30-180, retention time: 3-18 min).

CHAPTER IV

Discrimination of Complex Odors with Gas Chromatography-Mass Spectrometry Data by Texture Image Analysis and Machine Learning

4.1 Abstract

Conventional odor discrimination is generally performed by gas chromatography–mass spectrometry (GC–MS) that identifies specific marker molecules. Such marker identification process is, however, labor-intensive, and the limited number of identified marker molecules is often insufficient to discriminate complex odors. In this study, we have demonstrated a facile method for discriminating complex odors with GC–MS data by combining texture image analysis (TIA) and machine learning (ML). We extracted various texture features (i.e., contrast, energy, homogeneity, correlation, dissimilarity and angular second moment) of two-dimensional (2D) MS maps by TIA, and used them as datasets for ML. Each texture feature contains a lot of molecular information appeared in 2D MS maps, and thus serves as an effective parameter for discriminating complex odors. Based on this method, we successfully performed the discrimination of breath samples collected from the persons of different blood glucose levels with higher performances and reliability than the conventional approach

Keywords: *Odor discrimination, texture image analysis, machine learning, 2D MS map, GLCM*

4.2 Introduction

Odor analysis is a promising technique for non-invasively characterizing a subject based on the species and the concentrations of contained volatile chemical compounds. This type of analysis has recently attracted much attention in various scientific and industrial fields such as pathology,1-3) pharmacology,4,5) healthcare,6-12) food industry,13-15) fragrance and perfume industry,16-18) environmental conservation,19-21) agriculture22-24) and so on.25-27) The odor analysis is performed by two-step process consisting of i) a marker molecules identification and ii) a discrimination or classification of odors based on the identified specific markers. In odor analysis, GC–MS is conventionally employed for identifying marker molecules.28-33) However, the marker identification process is labor-intensive, and limited number of identified marker molecules is often insufficient to discriminate complex odors. The texture image analysis (TIA) is a useful way to effectively collect large amount of information in an image.34,35) TIA has recently been applied to medical image analyses and successfully demonstrated its performances on the tumor identification and the radiotherapy beyond the sense of human eyes.36,37) Despite the advantage of TIA, it has rarely been applied to odor discrimination. These backgrounds motivated us to investigate the applicability of TIA to the discrimination of complex odors.

In this study, we demonstrated the discrimination of human breath samples with GC–MS data by combining texture image analysis and machine learning (TIA–ML). In this method, various texture features were extracted from two-dimensional (2D) MS maps. Each texture feature contains a lot of molecular information appeared in 2D MS maps, and thus serves as an effective parameter for discriminating complex odors. Based on this method, we successfully performed the discrimination of breath samples collected from persons with different blood glucose levels. The performance and reliability of the TIA–

ML method were discussed in comparison with those of a conventional marker identification approach.

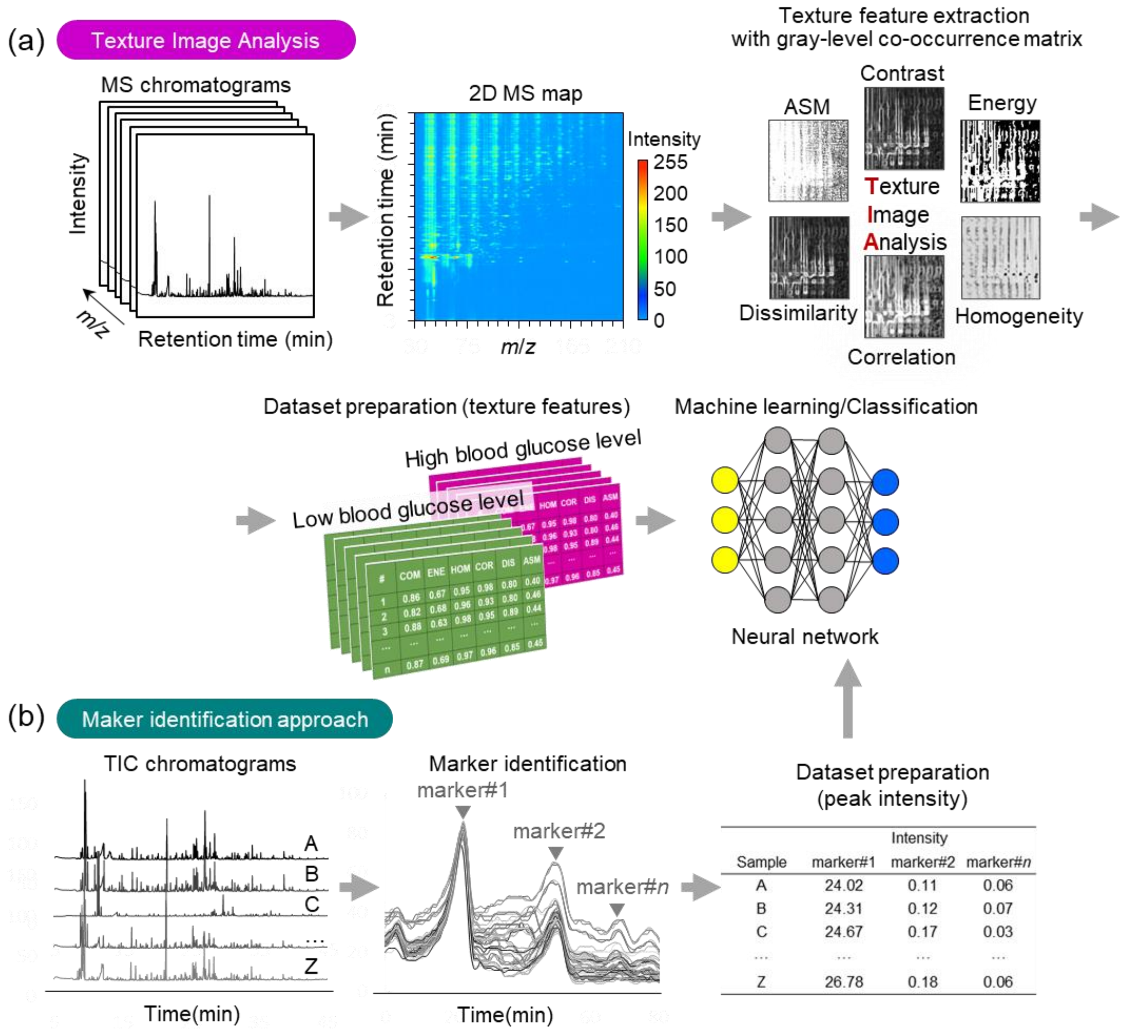
4.3 Experimental Section

Collections of Breath Samples and Blood Glucose Data. The human breath samples were collected from healthy volunteers under fasting condition (8–10 h). To control the blood glucose levels, the volunteers took a 150 mL aqueous solution of 50g glucose (TRELAN-G50, AY Pharmaceuticals). The blood glucose level of volunteers was measured by a glucometer with conventional fingerstick method and a flash glucose monitoring system (FreeStyle Libre, Abbott). Each 50 breath samples were collected from the persons with high blood glucose level (HBG, ⩾ 125 mg/dL) and low blood glucose level (LBG, < 120 mg/dL). The exhaled breath was collected using a 10 L gas sample bag (Smart bag PA, GL Sciences). The 500 mL of collected breath was then transferred to an adsorbent-filled tube (Packed Liner with Tenax GR, mesh 80/100 #2414-1021, GL Science Inc.) using an air pump at the pumping rate of 50 mL/min. The sample tubes were sealed and stored in a refrigerator at –18 °C until they were used for the GC–MS measurements.

Breath Component Analysis by GC–MS. Total ion current (TIC) chromatograms and MS chromatograms of the breath samples were obtained by GC–MS (Shimadzu, GCMS-QP2020) equipped with an inlet temperature control unit (OPTIC). A InertCap 5MS/NP capillary column (60 m length, 0.25 mm inner diameter, 1 μm thickness, GL Sciences) was used, and the temperature profile of GC oven was set as follows: (i) held at 40 °C for 5 min, (ii) elevating to 280 °C at a rate of 5 °C/min, and (iii) held at 280 °C for 5 min. The inlet temperature was set at 300 °C with split-less mode. The temperatures of the ion source and the GC-to-MS junction were both set at 200 °C. The vacuum pressure in the ionization chamber was 9.9 × 10−5 Pa. He (99.9999% pure) was used as a carrier gas in column and a purge gas, and the flow rates were set at 1 mL/min and 5 mL/min, respectively. The MS measurements were carried out with a single quadrupole MS analyzer in a mode of electron ionization with positive ion analysis and the full scan data acquisition. A mass to charge ratio (m/z) was characterized in the range of 35–300. The obtained data were analyzed by GCMS Solution ver. 4.45 SP1.

Texture Image Analysis and Machine Learning of 2D GC–MS Data. GC–MS data was analyzed by the TIA–ML method and the conventional marker identification approach. The workflows of the TIA–ML method and the conventional marker identification approach are shown in Figure.1(a) and (b). For the TIA–ML method, firstly, all MS chromatograms, i.e., the series of retention time–signal abundance data, were combined and converted into a 2D MS map as the functions of m/z (x-axis) and retention time (y-axis). The range of m/z and retention time used for analysis were 35–300 and 3–58 min, respectively. The image resolution of 2D MS map was set to be 1300 × 3700 pixels. The intensity of 2D MS map was scaled by a power law (□ = 0.5), displayed by 256 colors, and normalized via the highest peak using Matplotlib ver. 3.5.1. To investigate the robustness of the TIA-ML method, the influences of position alignment and noise reduction in 2D MS map were examined. The details of such image processing for 2D MS map can be seen in our previous study.38)

To extract texture features of the 2D MS map, TIA was performed with gray-level co-occurrence matrix (GLCM)39) using Scikit-image ver. 0.19.1. The GLCM functions characterize the textures of an image by calculating the number of pairs of pixels with specific values in a specified spatial distance, creating GLCM maps, and then extracting statistical texture feature values from the matrix of GLCM map. In this study, the distance of 1 pixel and the angle of 45° were used. GLCM maps of contrast, energy, homogeneity, correlation, dissimilarity, and angular second moment (ASM) were created from 2D MS



Sample	Intensity marker#1	 marker#2	 marker#*n*
A	24.02	0.11	0.06
B	24.31	0.12	0.07
C	24.67	0.17	0.03
...	...	...	...
Z	26.78	0.18	0.06

Figure.1 Graphical workflows of (a) texture image analysis (TIA) and (b) conventional marker identification approach for discriminating complex odors in GC–MS data.

map using the formulas shown in Table 1. Then the texture feature was obtained by a summation of the feature values of all pixels in a GLCM map. The extracted texture features were normalized and used as datasets for ML

For conventional marker identification approach, peaks were detected in TIC chromatograms and their intensities were used as datasets for ML. The peak detection was performed by CentWave method with the parameters of ppm=10, peak width minimum=1, peak width maximum=2, snthresh=100, mzdiff=6, prefilter scan number=0.01, prefilter scan abundance=3, and bw=100, which were optimized by using the method reported by Manier et al.40) In this study, for the simplicity, all detected peaks were used as the marker molecules for the discrimination of human breath samples, while the marker molecules are identified by carefully screening the detected peaks in the conventional odor discrimination study.

ML was conducted by a neural network algorithm. For ML, the datasets were divided into training data and testing data with a ratio of 70% and 30%, respectively. For enriching the training datasets while preventing overfitting, the data augmentation technique38) was employed. In this technique, the intensity of 2D MS maps was randomly modulated in the range of 0.0–10.0%. Consequently, the number of data increased by 100 datasets. The two-levels classification of breath samples (i.e., HBG and LBG) was performed with a multilayer perceptron, which is a class of feedforward artificial neural network, using Scikit-learn ver. 1.0.2. The classifiers were optimized by the hyper-parameters and operated with the parameters of hidden_layer_sizes = (128, 128), activation = 'relu', solver = 'adam', alpha = 1, max_iter = 1000 for the TIA-ML method and hidden_layer_sizes = (256, 512), activation = 'relu', solver = 'adam', alpha = 1, max_iter = 3000 for the conventional marker identification approach.

The odor discrimination performances in the TIA–ML method and the conventional marker identification approach were evaluated by calculating and comparing their classification accuracy, sensitivity, and specificity. The averaged area under the curve of receiver operating characteristic curve (AUC–ROC) was utilized to evaluate the reliability of classifier. The significance of each feature for the discrimination of human breath samples was evaluated with the p-value obtained in t-test.

Table 1 Texture features and formulas

Texture feature	Formula
Contrast	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$
Energy	$\sqrt{\sum_{i,j=0}^{N-1} (P_{ij})^2}$
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(1-j)^2}$
Correlation	$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-u)(j-u)}{\sqrt{(\sigma_i)^2(\sigma_j)^2}}\right]$
Dissimilarity	$\sum_{i,j=0}^{N-1} P_{i,j}\lvert i-j\rvert$
ASM (Angular second moment)	$\sum_{i,j=0}^{N-1} (P_{ij})^2$

4.4 Results and Discussion

2D MS Map and Texture Features of Human Breath Samples. Figure.2(a) and (b) show the representative TIC chromatograms and 2D MS maps of breath samples collected from the persons with two different blood glucose levels (i.e., HBG, ⩾ 125 mg/dL and LBG, < 120 mg/dL). TIC chromatogram is a primary form of GC–MS data used to identify marker molecules in the conventional approach. Each peak in the TIC chromatogram corresponds to a component molecule species in the tested breath, and each bright spot in the 2D MS map corresponds to a fragment peak of a component molecule species. Contrary to the TIC chromatograms, where the peak intensity can be quantitatively compared, the 2D MS maps were hardly distinguishable to the human eyes.

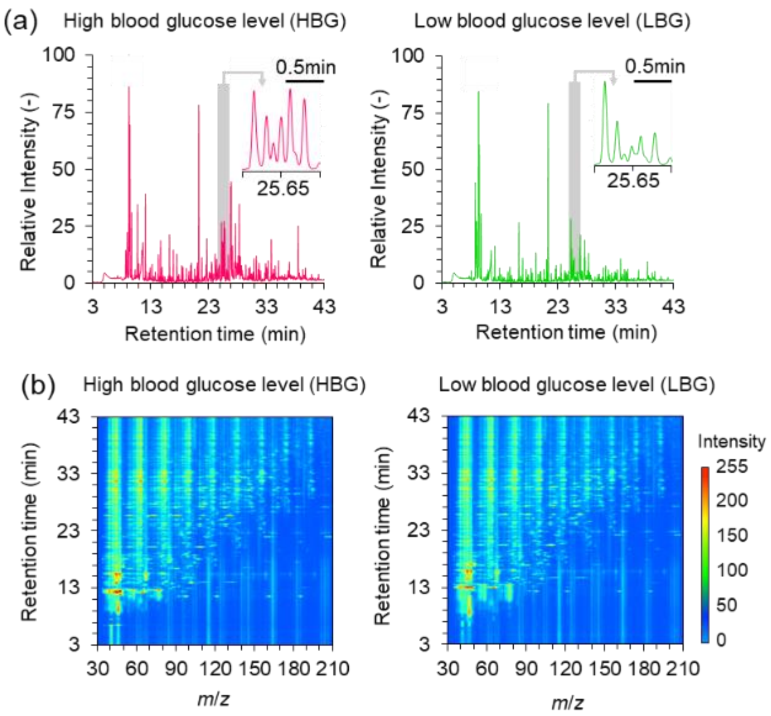


Figure.2 (a) TIC chromatograms and (b) 2D MS maps of breath samples collected from the persons with high blood glucose level (HBG, ⩾ 125 mg/dL) and low blood glucose level (LBG, < 120 mg/dL), respectively. For visibility, the 2D maps are shown in the restricted range (m/z: 30−210, retention time: 3−43 min).

Next, we extracted the features of GC–MS data. The texture features of the 2D MS maps were extracted by TIA with GLCM. Figure.3 shows the GLCM maps for the breath samples collected from the persons of HBG and LBG. Each texture feature was then obtained by a summation of feature values of all pixels in a GLCM map. The extracted texture features were normalized and used as datasets for ML.

We created a classifier by ML with a neural network algorithm. As a comparison, we also created a classifier by the conventional marker identification approach. For this purpose, we identified the peaks of marker molecules in the TIC chromatograms, and the peak

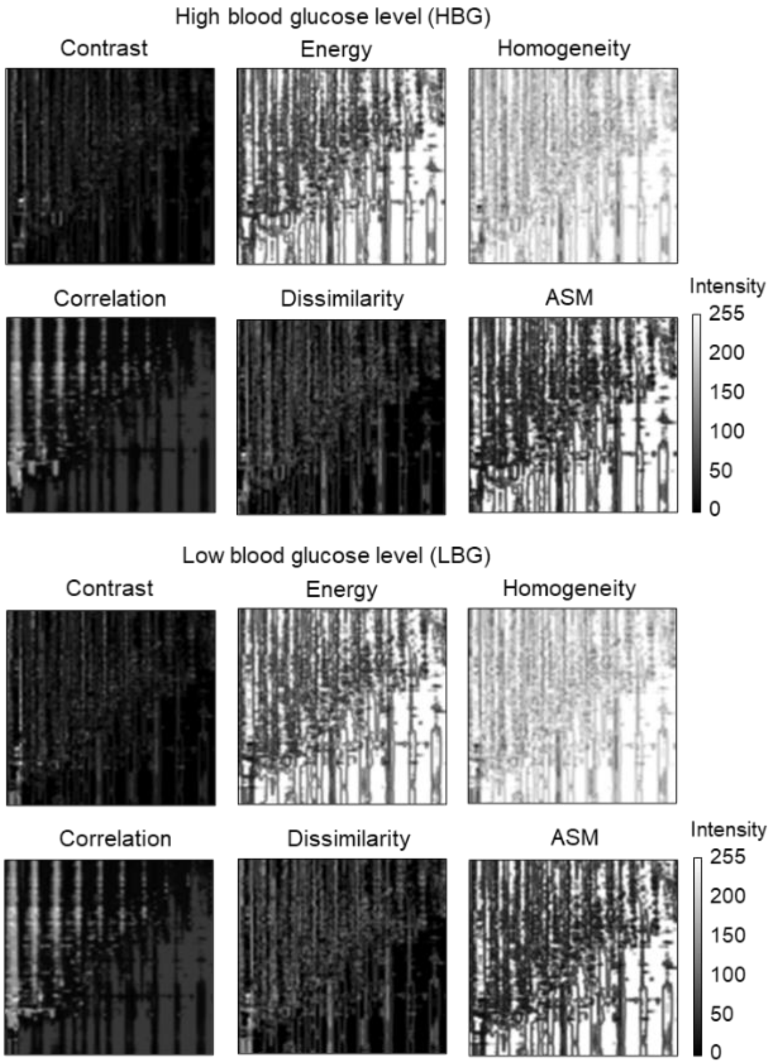


Figure.3 GLCM maps for texture features of contrast, energy, homogeneity, correlation, dissimilarity, and ASM.

intensities were used as datasets for ML. By using the classifiers, we calculated the classification accuracies for the test breath samples.

Classification Performance of TIA-ML Method for Human Breath Samples. Figure.4(a) shows the classification accuracy of breath samples of HBG and LBG, plotted as a function of the number of features employed for creating the classifier. The employed features were arranged in ascending order of the p-values. In the conventional marker identification approach, the classification accuracy was as low as 20.0% when employing a single feature. It tended to increase with increasing the number of employed features and reached to 100% with 50 features. On the other hand, in the TIA–ML method, the classification accuracy was 83.3% with a single feature, and reached to 100% with two features. These results clearly indicated that that the TIA–ML method provided a higher classification accuracy with fewer features than the conventional marker identification approach. Note that both of the specificity and sensitivity of the TIA–ML method reached to 100% with two features. Such excellent discrimination performance of the TIA–ML method can be interpreted by the fact that each texture feature contains a lot of molecular information.

Reliability of TIA-ML Method. To confirm the validity of above-mentioned classification performance, we evaluated the reliability of classifiers. Figure.4(b) shows the averaged AUC–ROC for the TIA–ML method and the conventional marker identification approach, plotted as a function of the number of features employed for creating the classifier. As well as the trends of classification accuracy in Figure.4(a), the AUC–ROC tended to increase by accompanying with the increase of the number of employed features. The AUC–ROC in the TIA–ML method reached to 1.00 with two features while that in the conventional marker identification approach was as low as 0.47. These results highlighted that the TIA–ML method showed better performances in both the accuracy and reliability for the discrimination of the human breath samples.

Advantage of Texture Feature. Here we discuss the contribution of each feature on the classification results in Figure.4(a). In the conventional marker identification approach, the classification accuracy decreased with increasing number of features due to a so-called

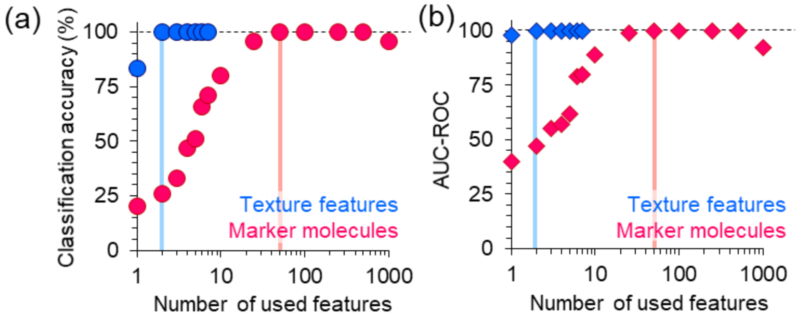


Figure.4 (a) The classification accuracy of the breath samples and (b) the averaged AUC-ROC for the TIA–ML method and the conventional marker identification approach, plotted as a function of the number of features employed for creating the classifier.

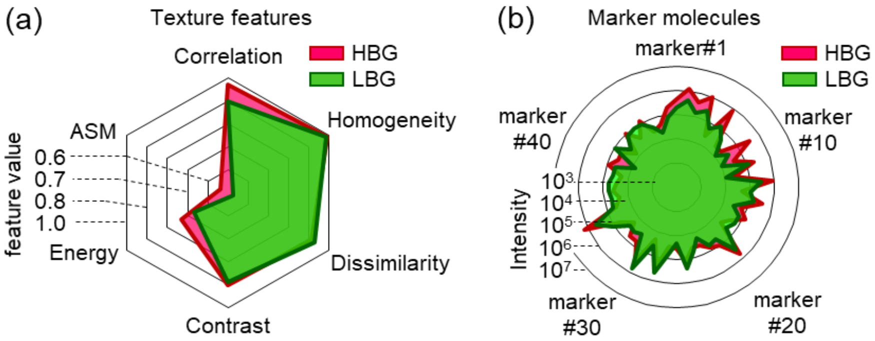


Figure.5 Radar charts for the feature values used in (a) the TIA and (b) the conventional marker identification approach, respectively. In these charts, the mean feature values of breath samples collected from the persons of HBG and LBG are plotted.

overlearning effect, in which the performance of classifier deteriorates by learning disturbing features. Interestingly, such an overlearning effect did not occur at all in the

TIA–ML method. This indicates that all texture features positively contributed to the classification.

To gain an in-depth understanding as to the role of extracted texture features, we quantitatively compared their feature values on the human breath samples of HBG and LBG. Figure.5(a) and (b) show the radar charts of the feature values for the TIA–ML method and the conventional marker identification approach, respectively. In these charts, the mean feature values are plotted. Note that the features are arranged in ascending order of the p-values and displayed clockwise on the charts. We found that there was no clear relationship between the substantial difference in the feature values and the arrangement order of the features in the TIA–ML method. A similar trend was also found in the chart of the conventional marker identification approach. Figure.6 shows the box-and-whisker plots for the texture features in the human breath samples of HBG and LBG. The results showed that the distributions of feature values more clearly separated for the features of earlier order. On the other hand, the distributions of feature values overlapped in some texture features such as dissimilarity, contrast, energy and ASM. Considering the fact that no overlapping occurred in either the classification accuracy or the AUC–ROC, the classification error in each texture feature could be negligible in the assembled features (i.e., combination of texture features). Such a beneficial role of texture feature might be available only when more maker molecules than contaminant molecules occupy the analyte odors.

Robustness of TIA–ML Method. Finally, we investigated the robustness of the TIA–ML method by examining the influences of position shift and noise in 2D MS maps. The position shift of spots occurs when a liquid phase of GC column deteriorates over time. The noise is caused by the deterioration of GC column and/or the contaminant molecules (i.e., non-marker molecules) in analytes. Figure.7 shows the box-and-whisker plots for representative texture features of the human breath samples of HBG and LBG when Figure.7 shows the box-and-whisker plots for representative texture features of the human breath

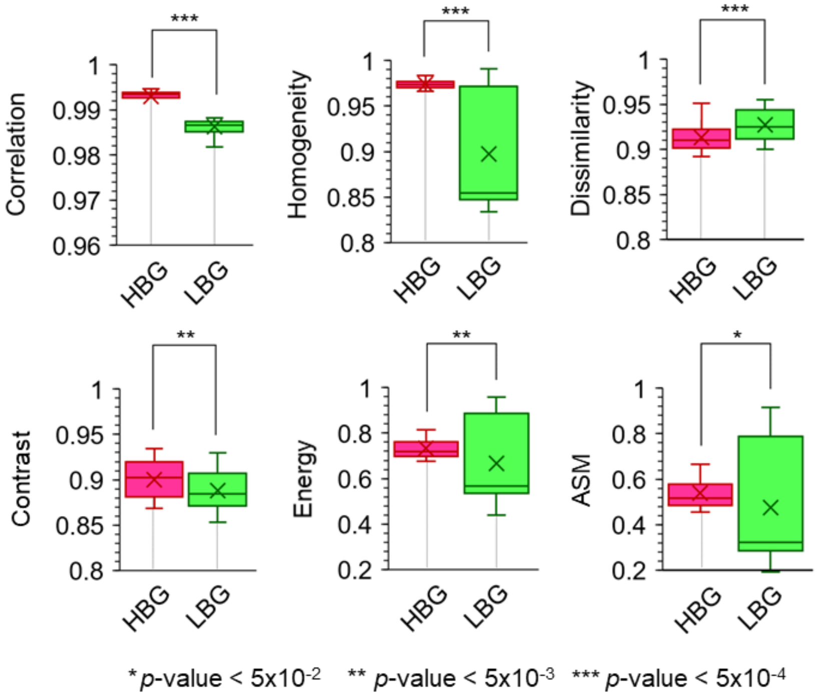


Figure.6 Box-and-whisker plots for the texture features of the breath samples collected from the persons of HBG and LBG.

samples of HBG and LBG when performing the image processing. As examples, the effects of position alignment and noise reduction in correlation and homogeneity are shown. The essential importance of position alignment and noise reduction was demonstrated in our previous study for the identification of marker molecules by image analysis.38) We found that the significance of each texture feature (i.e., p-value) was in the almost same order, regardless of the position alignment and/or the noise reduction. These results are reasonably interpreted by the principle of TIA, where the spatial relationship of pixels is emphasized in the texture features. It should be worth describing that the robustness to position shift and noise in TIC chromatograms is not available in the conventional marker identification approach. As such, the data analysis process in the TIA–ML method can be simpler than that in the conventional marker identification approach and thus the high-throughput discrimination of complex odors would be expected by the TIA–ML method.

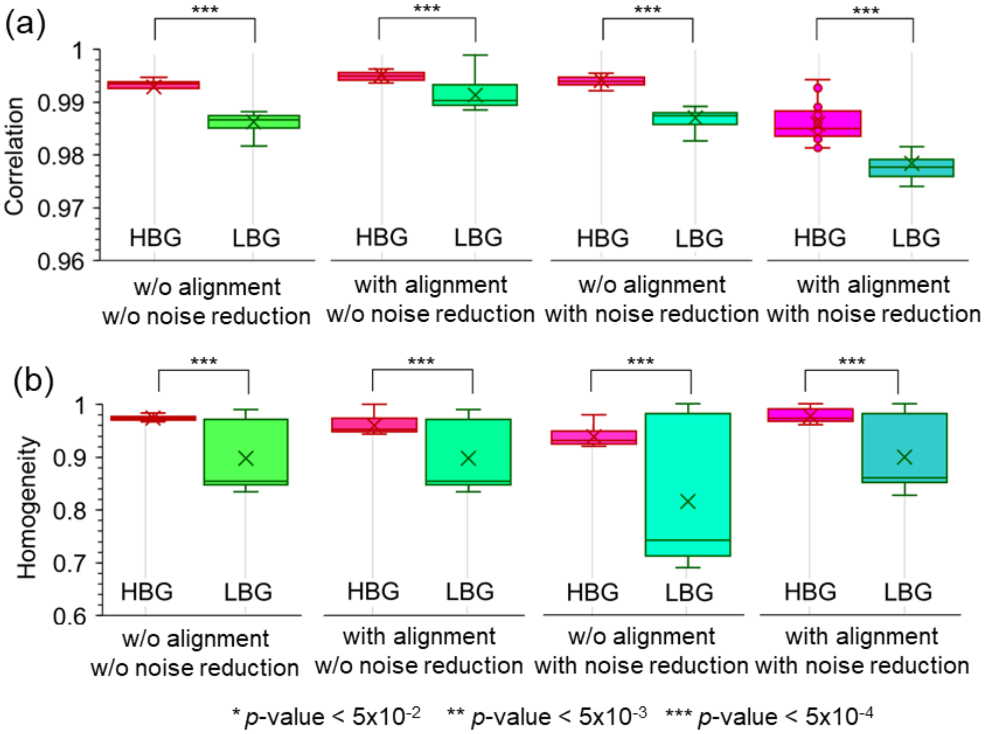


Figure.7 Box-and-whisker plots for the representative texture features (correlation and homogeneity) of the breath samples collected from the persons of HBG and LBG when performing the image processing (position alignment and noise reduction) with various combinations.

4.5 Conclusion

We demonstrated a facile method for discriminating complex odors with GC–MS data by combining texture image analysis and machine learning (i.e., TIA–ML method). In the proposed method, various texture features (i.e., contrast, energy, homogeneity, correlation, dissimilarity and ASM) of 2D MS maps were extracted by TIA with GLCM and used as datasets for ML. Contrary to the conventional marker identification approach, which relies on the limited number of marker molecules, each texture feature contains a lot of molecular information appeared in the 2D MS map, and thus served as an effective parameter for discriminating complex odors. By the TIA–ML method, we successfully performed the discrimination of breath samples collected from the persons of different blood glucose levels with higher performances and reliability than the conventional marker identification approach. While this study was limited to a two-levels classification, the TIA–ML method is essentially applicable to a multilevel classification. Thus, we believe that the TIA–ML method paves a novel avenue in complex odor analysis.

4.6 References

H. Haick et al., Chem. Soc. Rev., 43, 1423 (2014).

M. Hakim et al., Chem. Rev., 112, 5949 (2012).

M. M. Ioana et al., Medicina (Kaunas), 56, 118 (2020).

L. Schreiner et al., J. Nat. Prod., 83, 834 (2020).

H. Wakayama et al., Ind. Eng. Chem. Res., 58, 15036 (2019).

M. K. Nakhleh et al., ACS Nano, 11, 112 (2017).

F. Decrue et al., Anal. Chem., 93, 15579 (2021).

A. Z. Berna et al., ACS Infect. Dis., 7, 2596 (2021).

T. Güntner et al., ACS Sens., 4, 268 (2019).

G. Giovannini et al., ACS Sens., 4, 1408 (2021).

Z. Li et al., Anal. Chem., 93, 9158 (2021).

D. Maier et al., ACS Sens., 4, 2945 (2019).

E. Guichard et al., J. Agric. Food Chem., 68, 10318 (2020).

M. Bösl et al., J. Agric. Food Chem., 69, 1405 (2021).

J. E. Grimm et al., J. Agric. Food Chem., 67, 5838 (2019).

R. Toniolo et al., Anal. Chem., 85, 7241 (2013).

M. A. Teixeira et al., Ind. Eng. Chem. Res., 52, 963 (2013).

V. B. Xavier et al., Ind. Eng. Chem. Res., 59, 2145 (2020).

J. M. Estrada et al., Environ. Sci. Technol., 45, 1100 (2011).

J. Quintana et al., Environ. Sci. Technol., 50, 62 (2016).

G. Ašmonaitė et al., Environ. Sci. Technol., 52, 14381 (2018).

S. Giglio et al., Environ. Sci. Technol., 42, 8027 (2008).

Y. Wang et al., J. Agric. Food Chem., 53, 3563 (2005).

M. Bengtsson et al., J. Agric. Food Chem., 49, 3736 (2001).

Q. Wang et al., Sci. Rep., 10, 14856 (2020).

C. Wongchoosuk et al., Sensors, 9, 7234 (2009).

S. K. Jha, Rev. Anal. Chem., 36, 20160028 (2017).

M. P. Styczynski et al., Anal. Chem., 79, 966 (2007).

S. Zhang et al., ACS Omega, 5, 26402 (2020).

B. Demarcq et al., J. Agric. Food Chem., 69, 3175 (2021).

L. Sun et al., J. Agric. Food Chem., 69, 9350 (2021).

Z. Jia et al., ACS Omega, 3, 5131 (2018).

D. K. Trivedi et al., ACS Cent. Sci., 5, 599 (2019).

E. R. Thaler et al., Expert. Rev. Med. Devices., 2, 559 (2005).

N. Feizi et al., Trends. Analyt. Chem., 138, 116239 (2021).

N. B. Bahadure et al., Int. J. Biomed., 2017, 1 (2017).

T. S. Kumar et al., Biomed. Signal Process. Control., 73, 103440 (2022).

J. Chaiyanut et al., Anal. Chem., 93, 14708 (2021).

A. Kassner and R.E. Thornhill, Am. J. Neuroradiol., 31, 809 (2010).

S. K. Manier et al., Drug Test. Anal., 11, 752 (2019).

CHAPTER V

Human Breath Odor Based Individual Authentication by Integrated Sensor Array

5.1 Abstract

Exhaled breath contains thousand chemical compounds and serves as abundant information source to characterize the person. Here we demonstrate a primary study of breath odor sensing based individual authentication using artificial olfactory sensor array. The breath samples collected from 6 persons were tested by 16-channel chemiresistive sensor array and the sensing responses were analyzed by machine learning with random forest algorithm. The median accuracy of 96.4% was successfully achieved for the individual authentication for 6 persons. We found that the prediction accuracy and the reproducibility of individual authentication significantly improved by increasing the number of used sensors. Irrelevance between the error distribution of individual authentication and the feature score profiles of used sensor implies that discriminations of gender, age and nationality are of equal difficulty. These results provide an important foundation towards breath odor sensing based biometrics.

Keywords: *Breath, Chemiresitive sensor, Feature, Biometric.*

5.2 Introduction

Biometric authentication is a convenient and secure individual authentication method for cyber security in the information and communication technology (ICT) field. Its application range covers not only immigration control at airport but also access control of banking, personal computer (PC)/mobile phone and emerging intelligent vehicle (IVs).[1] To date, various techniques have been developed for biometric authentication, which include fingerprint/palmprint verification,[2] iris/retina recognition,[3] facial recognition,[4] hand and finger geometry,[5] voice biometry,[6] finger vein recognition[7] and ear acoustic authentication.[8] All these techniques solely rely on physical information, and thus have the risks of being unusable by information alternation with injury or being compromised by malicious information theft.

Human scent analysis/sensing is a new class biometric authentication technique using chemical information.[9-15] Since human scents such as exhaled breath and percutaneous gas have a strong genetic basis,[11,16,17] their chemical composition profiles are inherently different among individuals and therefore can potentially be utilized for individual authentication with low risks of information alternation/theft. Previously, human scent analysis/sensing based biometric authentication has been conceptualized and attempted mainly via percutaneous gas.[9-15] For example, Penn et al. analyzed the chemical component profiles of sweat odors from 197 adults using gas chromatograph-mass spectrometry (GC-MS) and identified 44 individual specific volatile organic compounds (VOCs).[10] Zheng et al. performed skin odor sensing by using artificial olfactory sensor array so-called electronic nose (e-nose) and classified the sensing data with 91.67 % of accuracy by machine learning.[13] Despite these previous achievements, the percutaneous gas sensing based individual authentication must have a limitation in its performance because the VOCs concentrations in percutaneous gas are usually lower (ppt to several tens ppb, ppt: parts per trillion, ppb: parts per billion) than the detection limit level of conventional chemical sensors and therefore the detectable number of VOCs species is restricted.[18] On the other hand, exhaled breath is known to have thousand VOCs and their concentrations are about three orders of magnitude higher than those of percutaneous gas (ppb to several ppm, ppm: parts per million).[18] In this respect, the breath odor sensing has a great potential to detect larger number of human-related VOCs species and achieve the higher performance in individual authentication compared with percutaneous gas sensing. However, the breath odor sensing has been mainly directed for pathology/disease diagnosis (e.g. cancer, diabetes, COVID-19),[19] and to best our knowledge, the feasibility of breath odor sensing based individual authentication has not been demonstrated so far.

5.3 Experimental Section

Breath odor sample preparation. The breath odor samples used in this study were collected from 6 healthy people (3 males and 3 females) with different nationalities (Thai, Chinese, Japanese) by using a 10 L gas sampling bag (Smart Bag PA CEK-10, GL Science Inc.). To exclude the influence of exogenous compounds originating from the diets and the tested environments, the breath samples were collected in the same room from the tested persons fasted for 6 h. The gas sampling bags filled with breath odor were stored for 12 h prior to the sensing measurements in order to stabilize the humidity condition inside the bag. For the breath component analysis, the gas sampling bag containing breath odor was connected to an adsorbent-filled sample tube (Packed Liner with Tenax GR, mesh 80/100 #2414-1021, GL Science Inc.) and 500 mL of breath odor was transferred to the sample tube with pumping at the rate of 50 mL/min. The sample tubes were sealed and stored in refrigerator at 4 ºC until conducting the gas chromatography–mass spectrometry (GC-MS) measurements.

Breath component analysis by GC-MS. Component analysis of the collected breath odor samples were conducted by GC-MS (Shimadzu, GCMS-QP2020) equipped with inlet temperature control unit (OPTIC4). For the GC-MS measurements, the collected chemical compounds in the sample tube were desorbed by rapidly increasing the injection port temperature to 300 ºC with split-less mode. The oven temperature was kept at 40 ºC for 5 min, then increased to 280 ºC at the rate of 5 ºC/min, and kept at 280 ºC for 5 min. The capillary column of InertCap 5MS/NP (60 m length, 0.25 mm inner diameter, 1 µm thickness, GL Science Inc.) was used to separate the desorbed compounds prior to MS analysis. The column flow rate and the purge flow rate of helium gas (99.9999% pure) were set to be 1 mL/min and 5 mL/min, respectively. Both the ion source temperature and the interface temperature of mass-spectrometer were fixed at 200 ºC during the measurements. The characterized mass to charge ratio (m/z) in the range of 35-300. The obtained data was analyzed by GCMS Solution ver. 4.45 SP1. The 2D MS maps and the 2D feature score maps were obtained by using NPFimg, i.e. the data analysis program we developed recently.[46]

Fabrication of artificial olfactory sensor array. 16 types of GC stationary phase material (GCM)-carbon black (CB) nanocomposite were prepared and used for sensing materials. The details of fabrication procedure and its usage can be seen elsewhere.[48,51-54] GCM-CB nanocomposites were made by mixing 10 mg carbon black (45μm Graphitized carbon black, Sigma) and 10 mg GC stationary phase materials (tetrahydroxyethylenediamine (THEED), GL Sciences/ N,N-Bis(2-cyanoethyl)formamide (BCEF), Tokyo Chemical Industry/LAC-3-R-728 (12% diethylene glycol succinate (DEGS), GL Sciences/DEGS, Supelco/ poly(ethylene succinate) (PES), Sigma/ UCON 75-H-90000, polyalkylene glycol (PAG) containing 75 wt% oxyethylene and 25 % oxypropylene groups, Shinwa Chemical Industries/1,2,3-Tris(2-cyanoethoxy)propane (TCEP), Supelco/SP-2330, poly (80% biscyanopropyl/20% cyanopropylphenyl siloxane), Supelco/SP-2340, poly (biscyanopropyl siloxane), Supelco/diglycerol, Tokyo Chemical Industry/Reoplex 400, GL Sciences/poly[di(ethylglycol)adipate] (PDEGA), Sigma/PEG4000, poly(ethyele glycol) 4000, Sigma/PEG20K, poly(ethyele glycol) 20000, United States Pharmacopeia (USP) Reference/PEG20M, poly(ethyele glycol) 20M, Shinwa Chemical Industries/free fatty acid phase (FFAP), Supelco) in 10 mL N,N-dimethylformamide (DMF, Wako). To uniformly disperse GCM and CB in solvent, the sonication was applied for 60 min at 38 kHz without addition of any dispersant. The as-prepared nanocomposite inks were deposited on an electrode-patterned Si substrate (n-type, with 100 nm-thick SiO_2 surface layer) to fabricate the 16-channel chemiresistive sensor array. Prior to the deposition, the comb-shaped Pt electrodes with Ti adhesive layer were first patterned on a 7 × 7 mm^2 sized substrate by photolithography and radio frequency (RF) sputtering. Gap distance and thickness of the electrodes were 40 μm and 400 nm, respectively. A SU-8 photoresist was then coated with 45 μm thickness on the electrode-patterned substrate by spin-coating and circular holes were made by photolithography. Each GCM-CB nanocomposite ink with 40 nL amount (40 shots at the rate of 1 nL/shot) was dropped at the circular holes by means of an ink-jet printing (custom-made, SIJ Technology Inc.). After depositing the GCM-CB nanocomposite inks, the device was annealed on a hotplate at 50 °C for 60 min and subsequently in a vacuum oven (100 Pa) at 50 °C for 60 min. The device was stored in the vacuum sealed bag until conducting the breath odor sensing measurements. The structural details of the fabricated sensor device were confirmed by an optical microscopy (OLYMPUS DP21).

Breath odor sensing measurement. The breath odor sensing data were collected by a homemade sensing module, which consists of a gas flow chamber, solenoid valves, an air pump and a sensor operation/data collection system. For the measurements, the flow of breath odor into the chamber was controlled by the pump at a rate of 100 mL/min. The sensing response was collected as a variation of output voltage by sequentially switching the flows of breath odor and N_2 carrier gas every 10 s with the solenoid valves. The sensing response was defined by the following equation: $\Delta V/V_0=(V-V_0)/V_0$, where V_0 and V are the output voltages under the flows of N_2 carrier gas and breath odor, respectively. All sensing measurements were performed at room temperature in air.

Data analysis with machine learning. Prior to the data analysis, the baseline correction was performed for the obtained sensing curves. The sensing responses ΔV were collected from 16-channel sensor array and used as dataset for machine learning. Totally 256 datasets of sensing response were obtained for each person (1536 sensing data for 6 people). For machine learning, random forest algorithm was employed to build classifiers. The models were optimized by the hyper-parameters and ran with 256 estimators (number of decision trees). A 9-fold cross-validation was used to confirm the reproducibility of classifier. The reliability of classifier was characterized by the average area under curve (AUC) of receiver operating characteristic (ROC) curve. The prediction accuracy and the coefficient of variation in prediction accuracy were computed to evaluate the performance of breath odor sensing based individual authentication.

5.4 Results and Discussion

In this study, we demonstrate a primary study for the breath odor sensing based individual authentication using artificial olfactory sensor array. In order to investigate the potential usage of breath odor for individual authentication, firstly we performed a GC-MS measurement and analyzed the individual specific molecular fragments. For the analysis, two-dimensional (2D) MS maps (m/z vs. retention time) were created and processed by using the recently developed data analysis program–NPFimg,[20] which combines an image processing and a machine learning. Figure. 1A-C show the 2D MS maps of breath odor samples collected from 3 persons (3 males). For the visibility, the 2D MS maps are shown in the restricted range (Full range 2D MS maps are shown in Figure. S1). Numerous molecular fragment signals are seen in the maps and many of them were common among the tested 3 persons. By learning the datasets of 2D MS maps, we succeeded in the individual authentication of 3 persons with 100 % of accuracy. Figure. 1D-F show the 2D feature score maps of molecular fragments contributed to discriminate the individual from the other two persons. Contrary to the 2D MS maps, the feature score maps were significantly different between the tested three persons. Note that the influence of exogenous compounds originating from the diets and the tested environment was negligible because the breath odor samples were simultaneously collected in the same environment from the persons who fasted for 6 h. We identified the individual-specific marker compounds, e.g. benzophenone, decanal, octane, tetradecane, undecane, which were consistently seen in the previous study of sweat odor based individual authentication (see details in Table S1).[10,12,15] Thus these results imply that each person has an original breath print derived from endogenous compounds and also indicate the potential feasibility of breath odor based individual authentication.

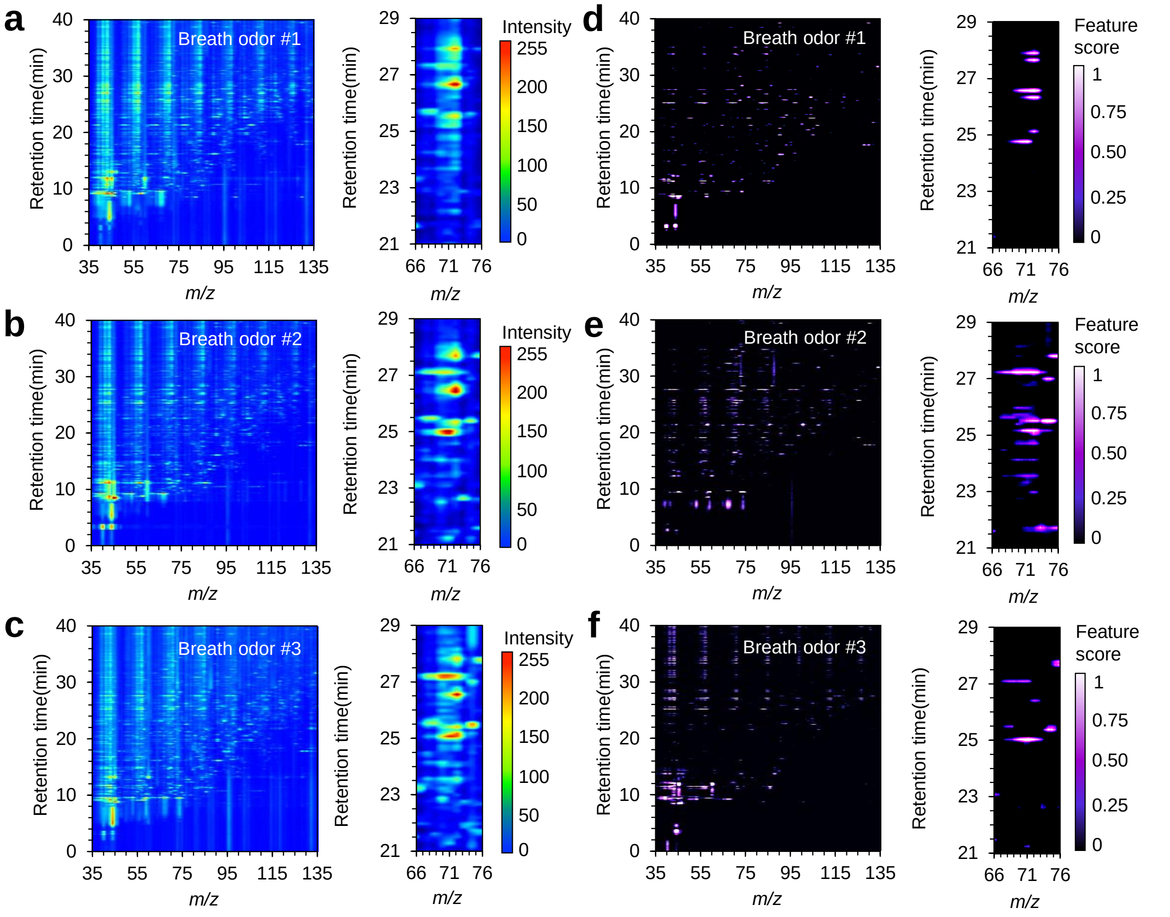


Figure. 1 (A-C) 2D MS maps and (D-F) 2D feature score maps of 3 tested persons (3 males) in wide rage view (left) and narrow range view (right). The 2D feature maps were obtained in comparison with the other two breath odor samples. The 2D feature maps were obtained in comparison with the other two breath odor samples.

We next examined the individual authentication via the breath odor sensing. The experimental workflow is shown in Figure. 2. The breath odor samples were first collected using a gas sampling bag (Figure. 2A). The collected breath odor sample was then flown into the sensing chamber installed with a 16-channel chemiresistive sensor array and the breath odor sensing was performed (Figure. 2B). The sensing materials used for the 16-channel sensor array are listed in Table S2. The sensing responses were acquired from the sensing curves of 16-channel sensors (Figure. 2C) and used as dataset for machine learning (Figure. 2D). We employed random forest algorithm for the machine learning (Figure. 2E) and demonstrated the individual authentication together with the feature profile evaluation of used sensors (Figure. 2F). The sensing was repeated 256 times for each person. We tested 6 persons (3 males, 3 females and ages 23-40) with various nationalities (Thai, Chinese and Japanese) as summarized in Table 1. Figure. 3A shows the five successive sensing curves obtained from 16-channel sensor array in the breath odor sensing of subject–V$^{\#1}$. The sensing characteristics such as the maximum sensing response, the initial sensing curve and the recovery curve were different between the sensors. These tendencies were also seen for the other subjects (V$^{\#2}$-V$^{\#6}$, Figure. S2-S6), while the sensing characteristics of each sensor strongly depended on the tested person. Figure. 3B shows the heatmaps of sensing responses of 16-channel sensor array for the tested 6 persons. The heatmaps were clearly different between the subjects. Such results are consistent with those of the GC-MS measurements and therefore anticipate the feasibility of breath odor sensing based individual authentication.

Table 1. The details of tested subjects for breath odor sensing based individual authentication.

Volunteer	Nationality	Age	Sex
$V^{\#1}$	Thai	23	Female
$V^{\#2}$	Thai	25	Male
$V^{\#3}$	Chinese	26	Female
$V^{\#4}$	Japanese	28	Male
$V^{\#5}$	Japanese	35	Male
$V^{\#6}$	Japanese	40	Female

Figure. 4A-E show the box-and-whisker plots of the accuracies of individual authentication for 2 persons, 3 persons, 4 persons, 5 persons and 6 persons, calculated by machine learning. The data are displayed as a function of the number of used sensors and the used sensors are arranged in the descending order of the coefficient of variation (CV) values in the sensing responses (Table S3). The median accuracy when using a single sensor were 92.8 %, 82.5 %, 64.7 %, 50.8 % and 30.4 % for individual authentications of 2 persons, 3 persons, 4 persons, 5 persons and 6 persons, respectively. The results indicate that the individual authentication tends to be difficult when the number of tested subjects increases. On the other hand, the accuracy of individual authentication was significantly improved when increasing the number of used sensors.

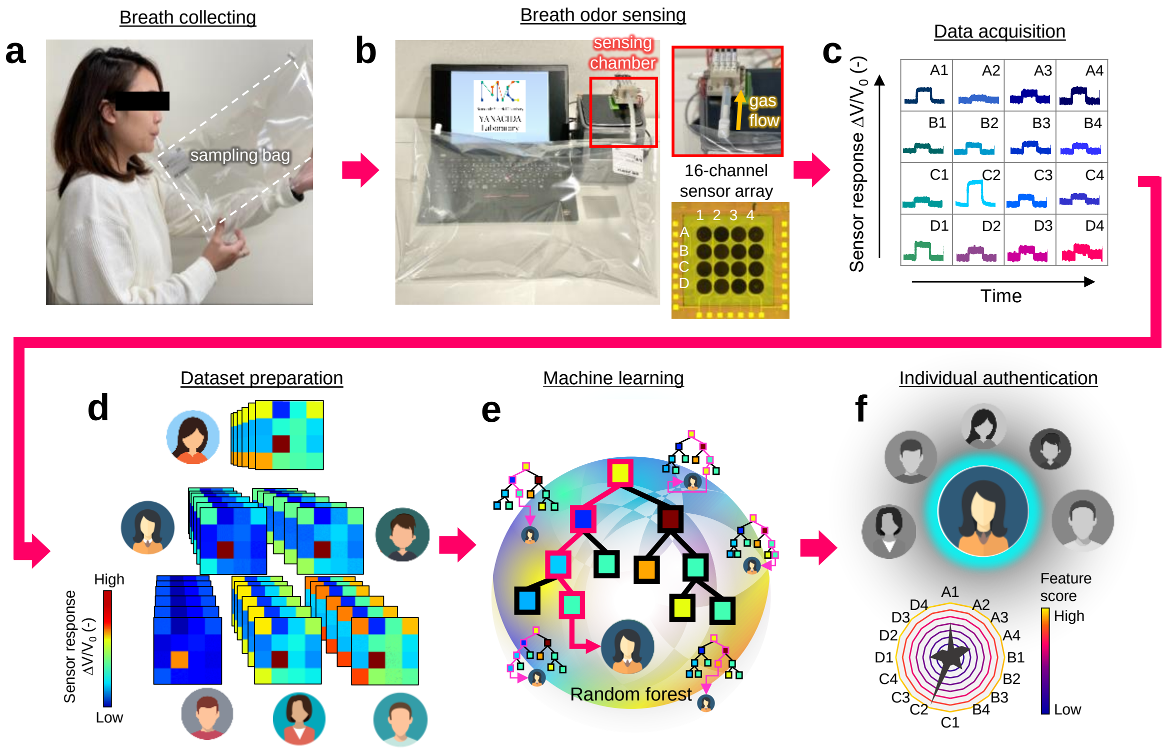


Figure. 2 Graphical workflow of breath odor sensing based individual authentication. (A) Breath odor sample collection using a gas sampling bag. (B) Breath odor sensing measurements using 16-channel sensor array. (C) Acquisition of sensing responses. (D) Dataset preparation for machine learning. (E) Machine learning with random forest algorithm. (F) Individual authentication and evaluation of feature profile of sensors.

The median accuracy for discriminating 6 persons successfully reached to 96.4 % by 16 sensors. The relationship between the number of subjects and the number of required sensors for individual authentication is displayed in Figure. 4F. The results indicate that a larger number of sensors are needed to discriminate complex odors, which is consistent with the claim in the recent review paper reported by Lee et al.[21] In other words, further discrimination of breath odors would be possible by increasing the number of used sensors. We next evaluated the reliability of the above breath odor sensing results. Figure.

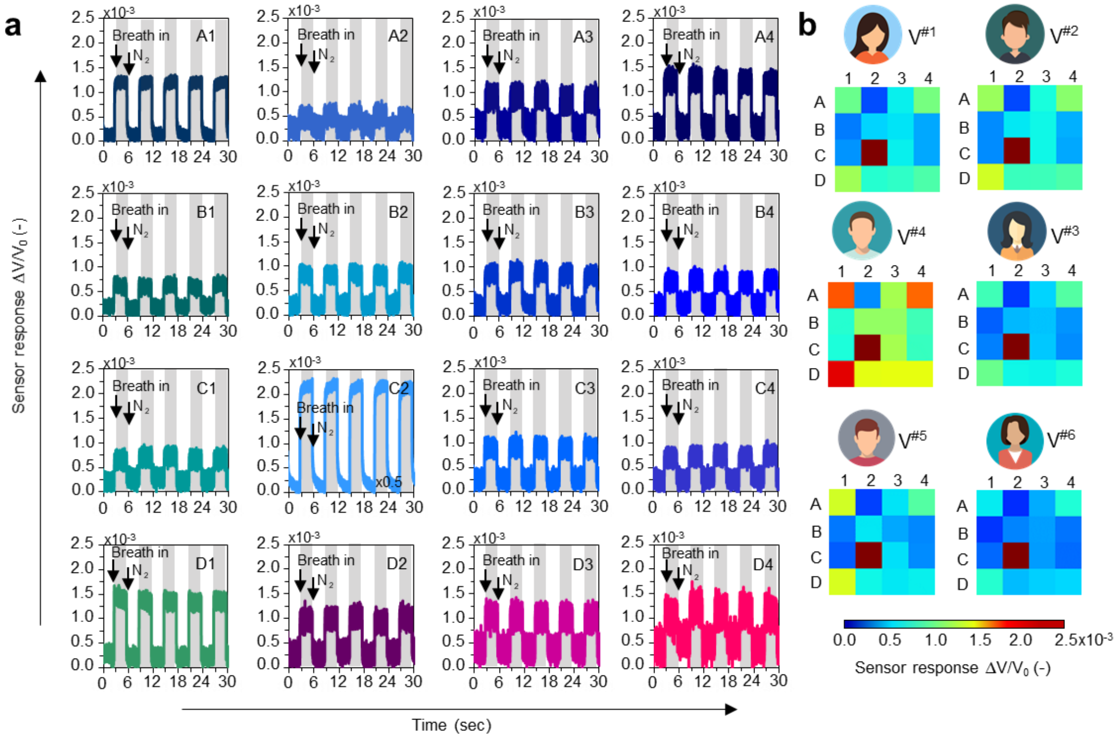


Figure. 3 (A) Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{\#1}$ after the baseline corrections. (B) Heatmaps of sensing responses of 16-channel sensor array for the breath odor sensing of each tested person (subject V$^{\#1}$-V$^{\#6}$).

4G and H show CV values for the accuracy of individual authentication and the averaged area under curve (AUC) of receiver operating characteristic (ROC) curve for the classifiers, which are presented as a function of used sensors. CV values in the accuracy significantly decreased and the averaged AUC of ROC curves increased as the number of used sensors increased. This shows that both the reproducibility of individual authentication and the reliability of classifiers also can be improved by using a larger number of sensors. All above results highlighted the potential feasibility of the breath odor sensing based individual authentication and the impact of number of integrated sensors on the performance of individual authentication.

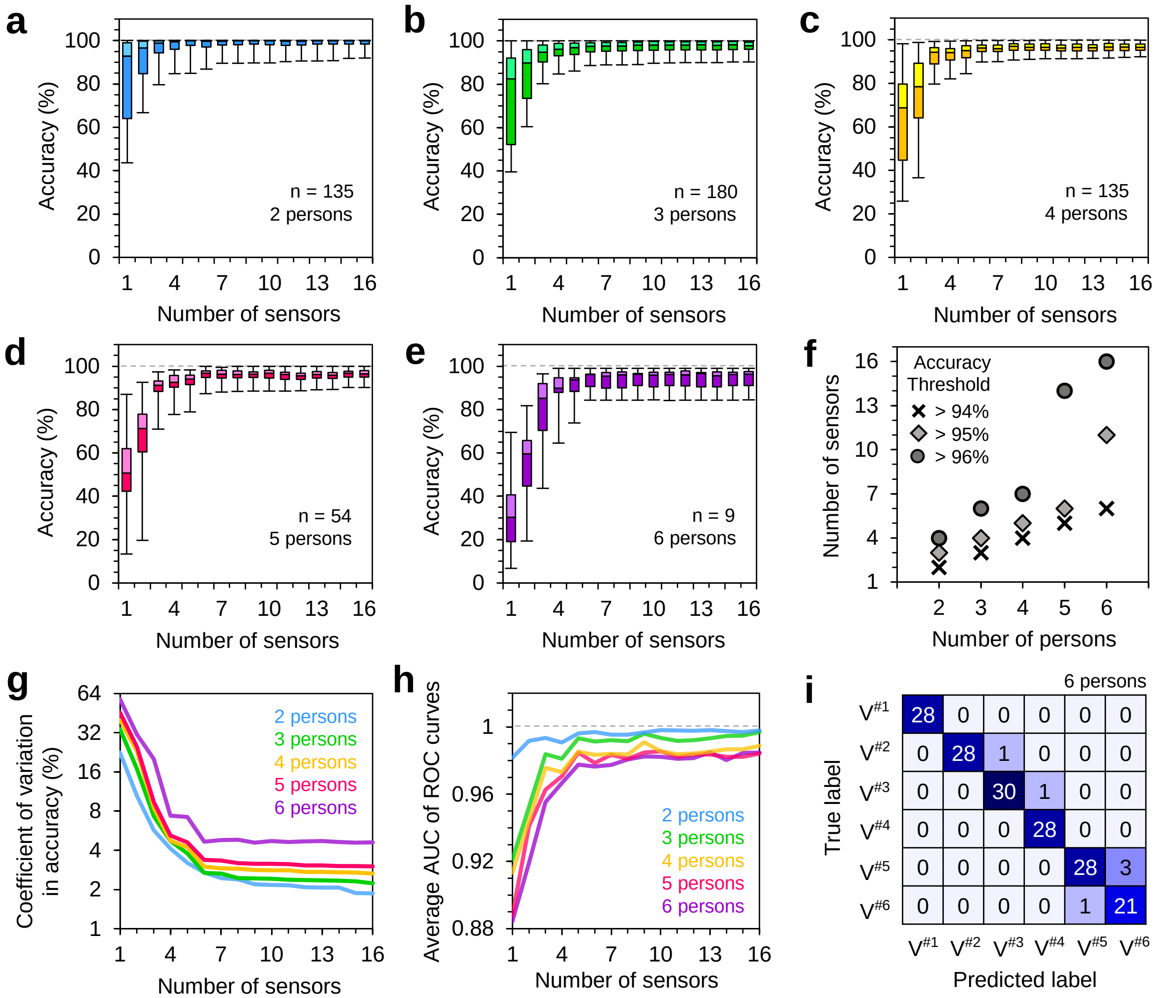


Figure. 4 Accuracy of breath odor sensing based individual authentication as a function of number of used sensors for (A) 2 persons, (B) 3 persons, (C) 4 persons, (D) 5 persons and (E) 6 persons, respectively. (F) A relationship between the number of persons and the number of required sensors with various threshold in accuracies (>94%, >95% and >96%). (G) Coefficient of variation in accuracy as a function of number of used sensors.

(H) Averaged AUC of ROC curves as a function of number of used sensors. (I) Confusion matrix for the breath odor sensing based individual authentication for 6 persons.

Here we discuss what critically determined the performance of breath odor sensing based individual authentication presented above. Figure. 4I shows the confusion matrix for the individual authentication of 6 persons. While the slight false identifications occurred, the errors were randomly distributed and their pattern was different in analytical batch. This result indicates that discrimination in gender, age and nationality is of equal difficulty. Figure. 5 shows the feature score profiles of used sensors for each tested subject. The data indicates that all sensors contributed to the individual authentication. Contrary to the result of confusion matrix, the gender-specific pattern was seen in the feature score profiles, implying that gender gives a great impact on the sensing response. This is reasonable by considering the difference in hormone balance between male and female as discussed in the previous reports.[14,15] Nevertheless, the individual authentication result and the feature score profile was irrelevant. This suggested that the false identification in our study might be caused by the fluctuation/instability of sensor responses and the performance of individual authentication would be better by improving the robustness of sensing materials.[22]

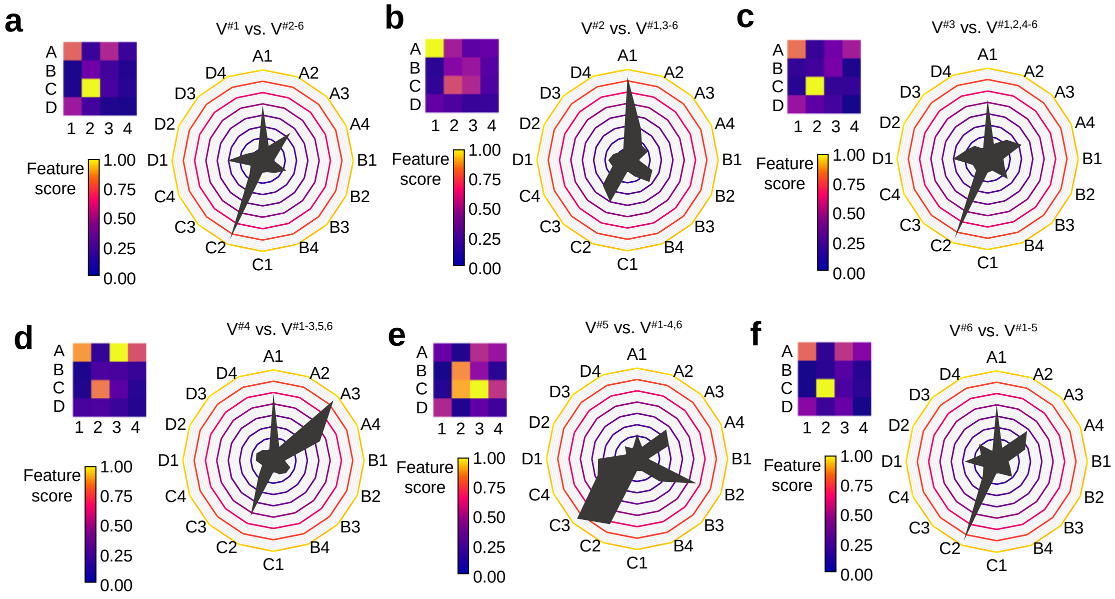


Figure. 5 Feature score patterns of 16-channel sensor array (heatmaps and radar charts) for (A) $V^{\#1}$ vs. $V^{\#2\text{-}\#6}$, (B) $V^{\#2}$ vs. $V^{\#1,\#3\text{-}\#6}$, (C) $V^{\#3}$ vs. $V^{\#1,\#2,\#4\text{-}\#6}$, (D) $V^{\#4}$ vs. $V^{\#1\text{-}\#3,\#5,\#6}$, (E) $V^{\#5}$ vs. $V^{\#1\text{-}\#4,\#6}$, (F) $V^{\#6}$ vs. $V^{\#1\text{-}\#5}$, respectively.

5.5 Conclusion

In conclusion, we demonstrated the primary study of breath odor sensing based individual authentication using artificial olfactory sensor array. The breath odor samples collected from 6 people were tested by 16-channel chemiresistive sensor array and the acquired sensing responses were analyzed by machine learning with random forest algorithm. The median accuracy of 96.4 % was successfully achieved for the individual authentication of 6 persons. We found that the accuracy and the reproducibility significantly improved by increasing the number of used sensors. While the breath odor sensing based individual authentication was demonstrated for the fasted subjects in this study, it still remains a challenging issue to demonstrate its feasibility under the interferences of disease related metabolites and exogenous compounds originating from the diets and the tested environment towards the practical application.[23] The barrier must be overcome by utilizing a larger number of sensors and extracting a larger number of features from the sensing curves. We believe that our findings in this study provide an important foundation towards breath odor sensing based biometrics.

5.6 References

1. K. Jain, Nature 2007, 449, 38-40.
2. W. Yang, S. Wang, N. M. Sahri, N. M. Karie, M. Ahmed and C. Valli, Sensors 2021, 21, 6163; J. J. Lozoya-Santos, V. Sepúlveda-Arróniz, J. C. Tudon-Martinez, R. A. Ramirez-Mendoza, Cogn. Syst. Res. 2019, 55, 175-191; A. A. Elngar and M. Kayed, Open Comput. Sci. J. 2020, 10, 17-29.
3. L. Hong, Y. Wan and A. Jain, IEEE Trans. PAMI 1998, 20, 777-789; W. Jia, D.-S. Huang and D. Zhang, Pattern Recognit. 2008, 41, 1504-1513.
4. J. Daugman, IEEE Trans. CSVT 2004, 14, 21-30; H. Borgen, P. Bours and S. D. Wolthusen, IEEE Computer Soc. 2008, 1056–1062.
5. Mian, M. Bennamoun and R. Owens, IEEE Trans. PAMI 2007, 29, 1927-1943.
6. Kumar and Y. Zhou, IEEE Trans. Image Process. 2012, 21, 2228-2244.
7. J. Yang, J. Chen, Y. Su, Q. Jing, Z. Li, F. Yi, X. Wen, Z. Wang, Z. L. Wang, Adv. Mater. 2015, 27, 1316-1326.
8. N. Miura, A. Nagasaka and T. Miyatake. Mach. Vis. Appl. 2004, 15, 194-203.
9. T. Arakawa, T. Koshinaka, S. Yano, H. Irisawa, R. Miyahara and H. Imaoka, 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp. 1-4.
10. Wongchoosuk, M. Lutz and T. Kerdcharoen, Sensors 2009, 9, 7234-7249; I. Rodriguez-Lujan, G. Bailador, C. Sanchez-Avila, A. Herrero, G. Vidal-de-Miguel, Knowl. Based Syst. 2013, 52, 279–289.
11. P. Inbavalli and G. Nandhini, Int. J. Comput. Sci. Inf. Technol. Adv. Res. 2014, 5, 6270-6274; S. K. Jha, Rev. Anal. Chem. 2017, 36, 20160028; M. C. Prieto-Blanco, S. P. Barba, Y. Moliner-Martínez, P. Campíns-Falcó, J. Chromatogr. A 2019, 1596, 241–249.

12. J. Penn, E. Oberzaucher, K. Grammer, G. Fischer, H. A. Soini, D. Wiesler, M. V. Novotny, S. J. Dixon, Y. Xu and R. G. Brereton, J. R. Soc. Interface 2007, 4, 331–340.
13. B. Holbert, H. P. Whitelam, L. J. Sooter, L. A. Hornak and J. M. Dawson, Netw. Model. Anal. Health Inform. Bioinform. 2015, 4, 22.
14. S. K. Jha and K. Hayashi, Int. J. Mass Spectrom. 2017, 415, 92-102.
15. Y. Zheng, H. Li, W. Shen and J. Jian, Sens. Actuators A 2019, 285, 395–405.
16. A. Sabilla, Z. A. Cahyaningtyas, R. Sarno, A. Al Fauzi, D. R. Wijaya and R. Gunawan, 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), 2021, pp. 109-115
17. P. Kanakam, A. S. N. Chakravarthy, Soft Computing 2021, 25, 13015–13025.
18. Lenochova P., Havlicek J. (2008) Human Body Odour Individuality. In: Hurst J.L., Beynon R.J., Roberts S.C., Wyatt T.D. (eds) Chemical Signals in Vertebrates 11. Springer, New York, NY.
19. Pol, G. H. Renkema, A. Tangerman, E. G. Winkel, U. F. Engelke, A. P. M. de Brouwer, K. C. Lloyd, R. S. Araiza, L. van der Heuvel, H. Omran, H. Olbrich, M. O. Elberink, C. Gilissen, R. J. Rodenburg, J. O. Sass, K. O. Schwab, H. Schäfer, H. Venselaar, J. S. Sequeira, H. J. M. Op den Camp and R. A. Wevers, Nat. Genet. 2018, 50, 120-129.
20. Toma, S. Suzuki, T. Arakawa, Y. Iwasaki and K. Mitsubayashi, Sci. Rep. 2021, 11, 10415.
21. J.-W. Yoon and J.-H. Lee, Lab Chip 2017, 17, 3537-3557; K. Inada, H. Kojima, Y. Cho-Isoda, R. Tamura, G. Imamura, K. Minami, T. Nemoto and G. Yoshikawa, Sensors 2021, 21, 4742.

22. M. Leja, J. M. Kortelainen, I. Polaka, E. Turppa, J. Mitrovics, M. Padilla, P. Mochalski, G. Shuster, R. Pohle, D. Kashanin, R. Klemm, V. Ikonen, L. Mezmale, Y. Y. Broza, G. Shani and H. Haick, Cancer 2021, 127, 1286-1292; G. Konvalina and H. Haick, Acc. Chem. Res. 2014, 47, 66-76.
23. J. Shin, S.-J. Choi, I. Lee, D.-Y. Youn, C. O. Park, J.-H. Lee, H. L. Tuller and I.-D. Kim, Adv. Funct. Mater. 2013, 23, 2357-2367.
24. R. Xing, L. Xu, J. Song, C. Zhou, Q. Li, D. Liu and H. W. Song, Sci. Rep. 2015, 5, 10717.
25. B. Shan, Y. Y. Broza, W. Li, Y. Wang, S. Wu, Z. Liu, J. Wang, S. Gui, L. Wang, Z. Zhang, W. Liu, S. Zhou, W. Jin, Q. Zhang, D. Hu, L. Lin, Q. Zhang, W. Li, J. Wang, H. Liu, Y. Pan and H. Haick, ACS Nano 2020, 14, 9, 12125–12132.
26. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, W. Tanaka, B. Samransuksamer, G. Zhang, J. Liu, M. Kanai and T. Yanagida, Anal. Chem. 2021, 93, 14708-14715.
27. S.-Y. Jeong, J.-S. Kim and J.-H. Lee, Adv. Mater. 2020, 32, 2002075.
28. W. Li, K. Nagashima, T. Hosomi, C. Wang, Y. Hanai, A. Nakao, A. Shunori, J. Liu, G. Zhang, T. Takahashi, W. Tanaka, M. Kanai and T. Yanagida, ACS Sens. 2021, in press.
29. E. Belizário, J. Faintuch and M. G. Malpartida, Front. Cell. Infect. Microbiol. 2021, 10, 564194; J. D. Pleil, M. A. Stiegel and T. H. Risby, J. Breath Res. 2013, 7, 017107.

5.7 Supporting Information

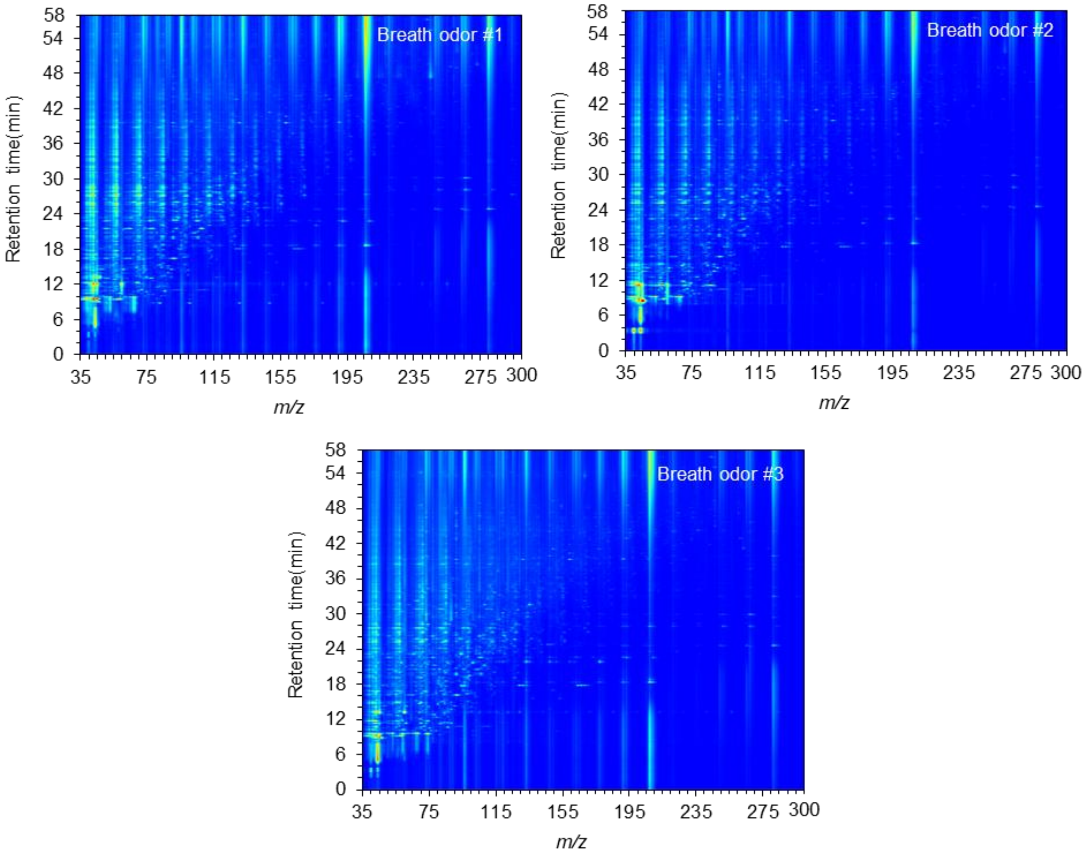
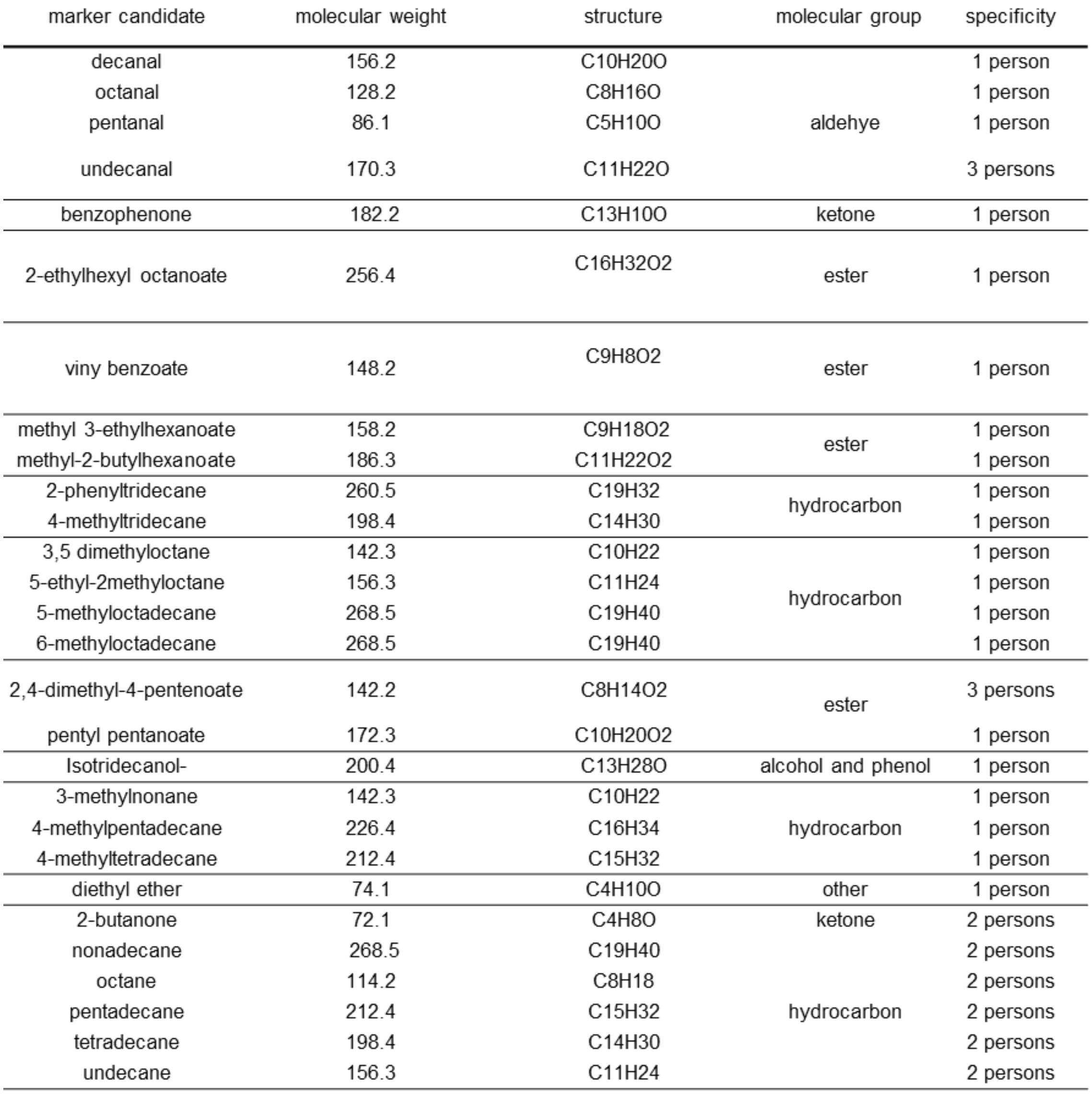


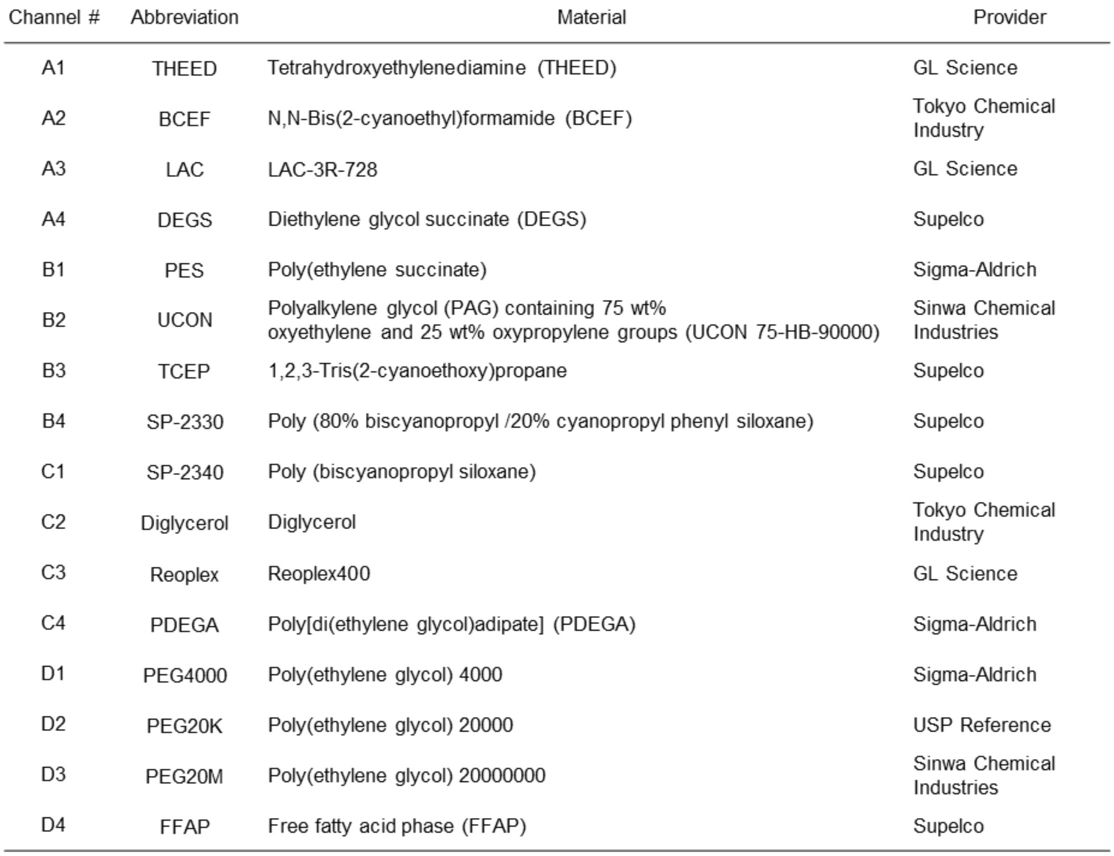
Figure S1. Full range 2D MS maps for the tested 3 persons (3 males), created by *NPFimg*.

Table S1. List of specific marker compounds for individual authentication



marker candidate	molecular weight	structure	molecular group	specificity
decanal	156.2	C10H20O	aldehye	1 person
octanal	128.2	C8H16O		1 person
pentanal	86.1	C5H10O		1 person
undecanal	170.3	C11H22O		3 persons
benzophenone	182.2	C13H10O	ketone	1 person
2-ethylhexyl octanoate	256.4	C16H32O2	ester	1 person
viny benzoate	148.2	C9H8O2	ester	1 person
methyl 3-ethylhexanoate	158.2	C9H18O2	ester	1 person
methyl-2-butylhexanoate	186.3	C11H22O2		1 person
2-phenyltridecane	260.5	C19H32	hydrocarbon	1 person
4-methyltridecane	198.4	C14H30		1 person
3,5 dimethyloctane	142.3	C10H22	hydrocarbon	1 person
5-ethyl-2methyloctane	156.3	C11H24		1 person
5-methyloctadecane	268.5	C19H40		1 person
6-methyloctadecane	268.5	C19H40		1 person
2,4-dimethyl-4-pentenoate	142.2	C8H14O2	ester	3 persons
pentyl pentanoate	172.3	C10H20O2		1 person
Isotridecanol-	200.4	C13H28O	alcohol and phenol	1 person
3-methylnonane	142.3	C10H22	hydrocarbon	1 person
4-methylpentadecane	226.4	C16H34		1 person
4-methyltetradecane	212.4	C15H32		1 person
diethyl ether	74.1	C4H10O	other	1 person
2-butanone	72.1	C4H8O	ketone	2 persons
nonadecane	268.5	C19H40	hydrocarbon	2 persons
octane	114.2	C8H18		2 persons
pentadecane	212.4	C15H32		2 persons
tetradecane	198.4	C14H30		2 persons
undecane	156.3	C11H24		2 persons

Table S2. The list of sensing materials used for the 16-channel sensor array.



Channel #	Abbreviation	Material	Provider
A1	THEED	Tetrahydroxyethylenediamine (THEED)	GL Science
A2	BCEF	N,N-Bis(2-cyanoethyl)formamide (BCEF)	Tokyo Chemical Industry
A3	LAC	LAC-3R-728	GL Science
A4	DEGS	Diethylene glycol succinate (DEGS)	Supelco
B1	PES	Poly(ethylene succinate)	Sigma-Aldrich
B2	UCON	Polyalkylene glycol (PAG) containing 75 wt% oxyethylene and 25 wt% oxypropylene groups (UCON 75-HB-90000)	Sinwa Chemical Industries
B3	TCEP	1,2,3-Tris(2-cyanoethoxy)propane	Supelco
B4	SP-2330	Poly (80% biscyanopropyl /20% cyanopropyl phenyl siloxane)	Supelco
C1	SP-2340	Poly (biscyanopropyl siloxane)	Supelco
C2	Diglycerol	Diglycerol	Tokyo Chemical Industry
C3	Reoplex	Reoplex400	GL Science
C4	PDEGA	Poly[di(ethylene glycol)adipate] (PDEGA)	Sigma-Aldrich
D1	PEG4000	Poly(ethylene glycol) 4000	Sigma-Aldrich
D2	PEG20K	Poly(ethylene glycol) 20000	USP Reference
D3	PEG20M	Poly(ethylene glycol) 20000000	Sinwa Chemical Industries
D4	FFAP	Free fatty acid phase (FFAP)	Supelco

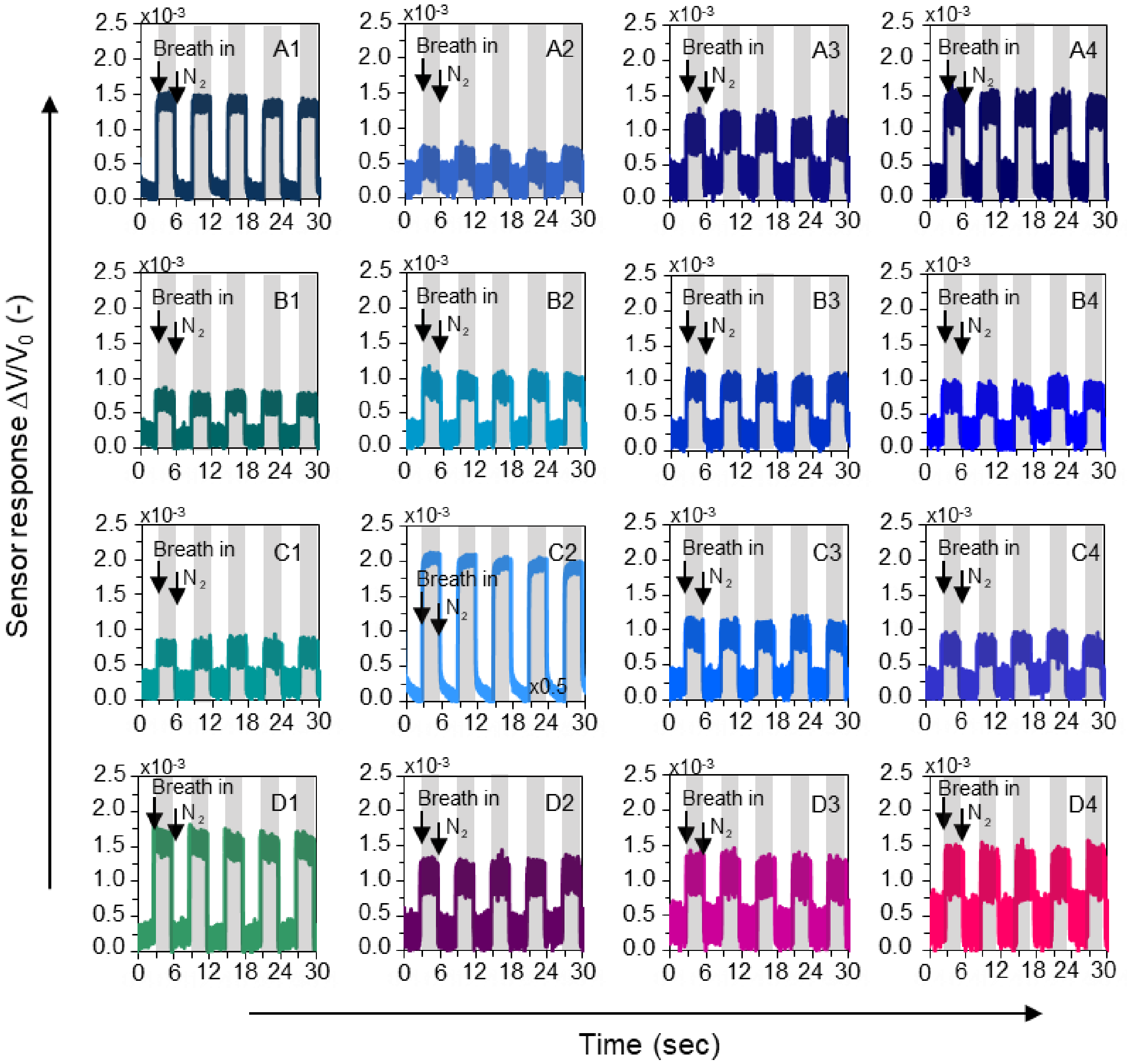


Figure S2. Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{\#2}$ after the baseline corrections.

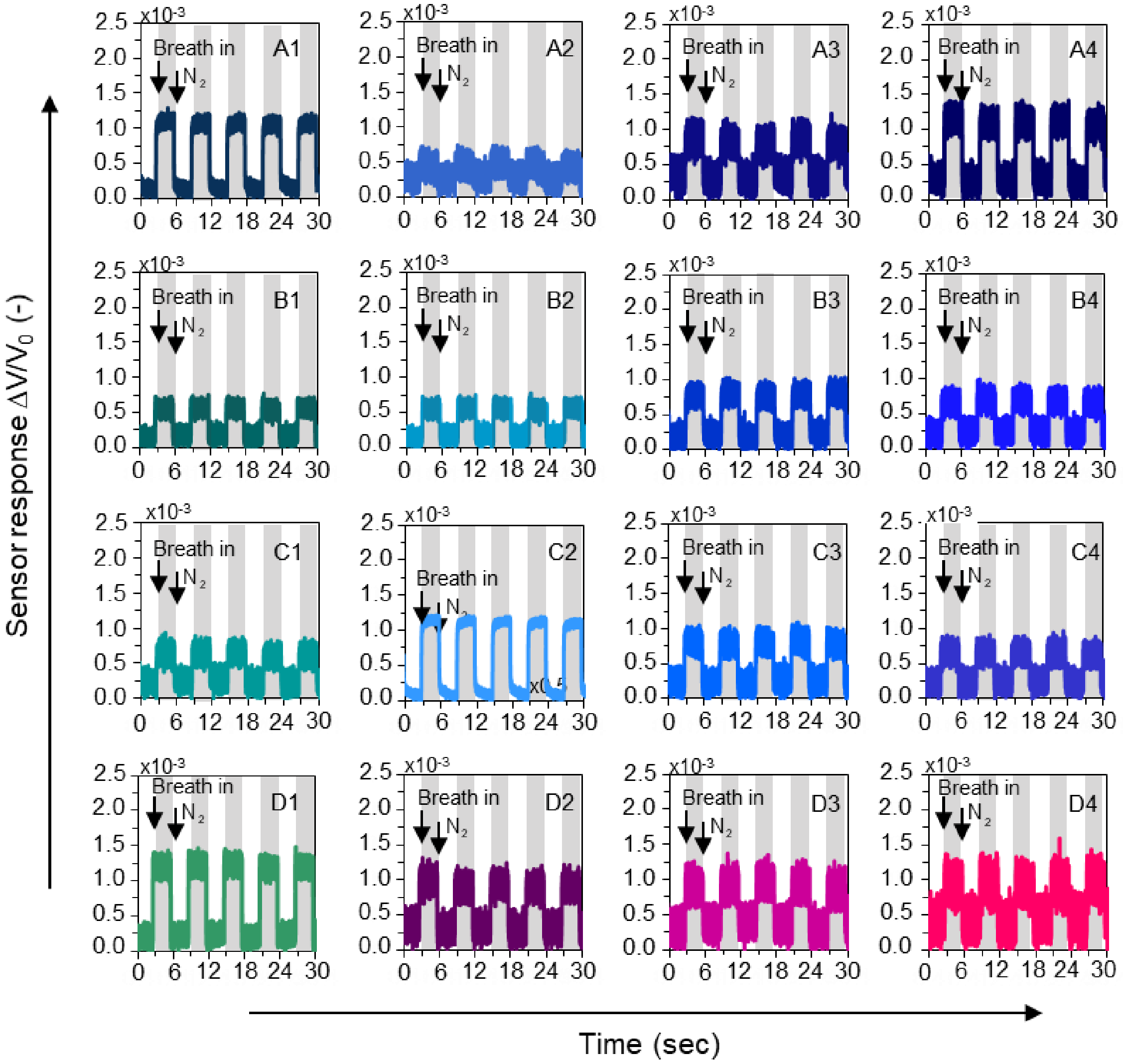


Figure S3. Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{\#3}$ after the baseline corrections.

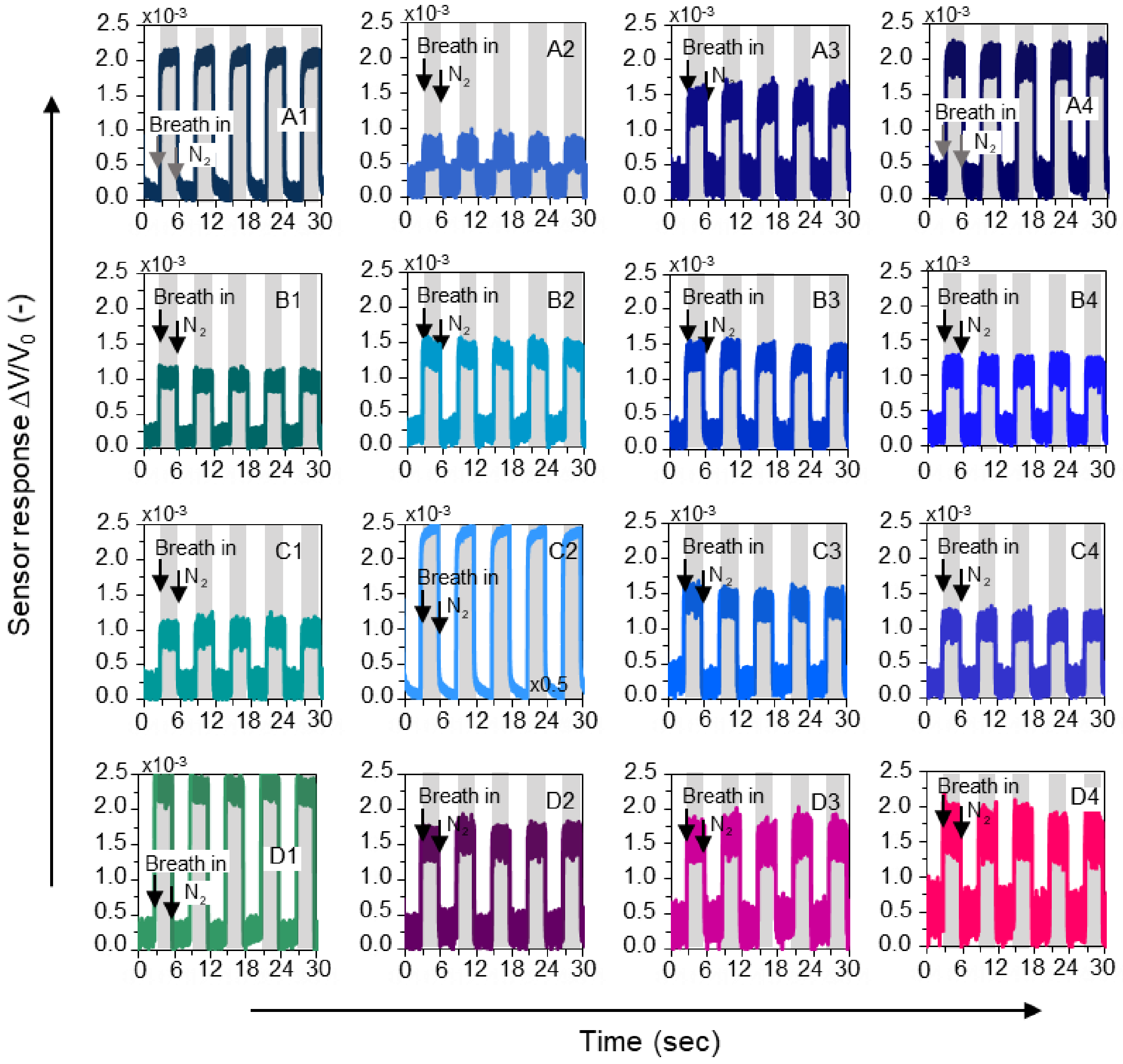


Figure S4. Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{\#4}$ after the baseline corrections.

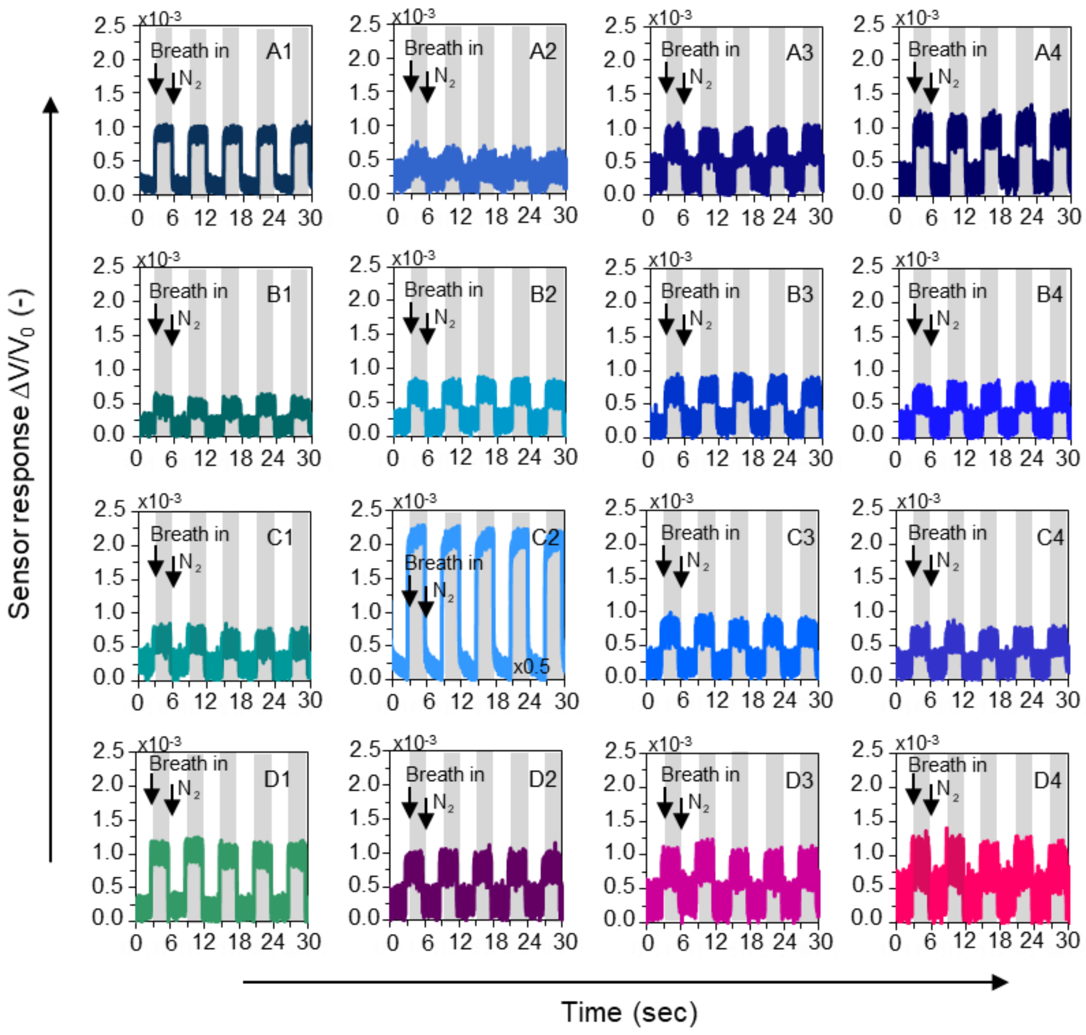


Figure S5. Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{#5}$ after the baseline corrections.

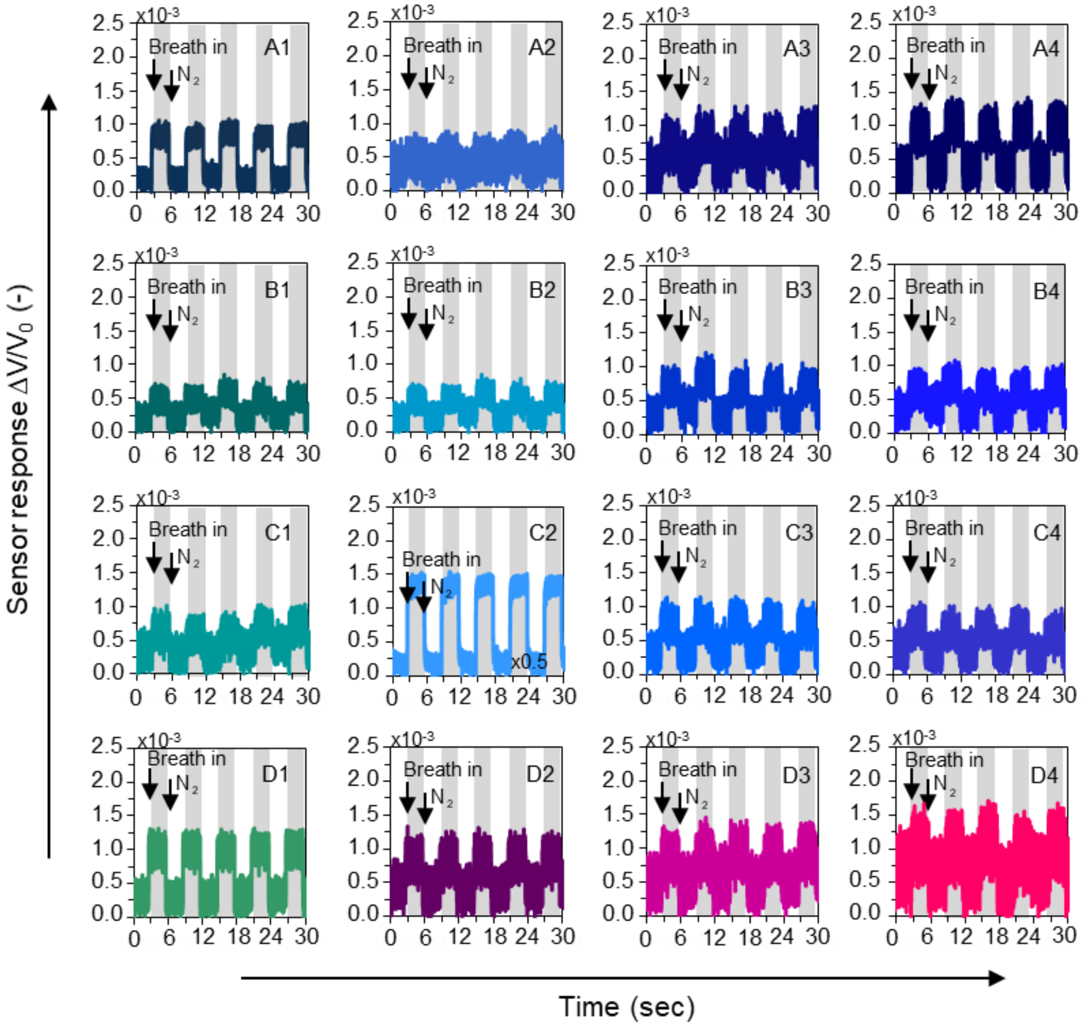
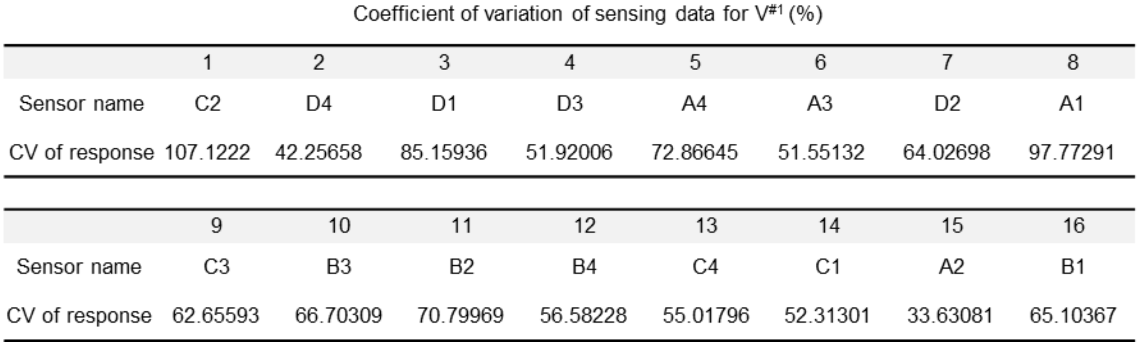


Figure S6. Sensing curves of 16-channel sensor array for the breath odor sensing of subject-V$^{\#6}$ after the baseline corrections.

Table S3. Coefficient of variation (CV) values in the sensing responses of used sensors.



Coefficient of variation of sensing data for $V^{\#1}$ (%)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sensor name	C2	D4	D1	D3	A4	A3	D2	A1	C3	B3	B2	B4	C4	C1	A2	B1
CV of response	107.1222	42.25658	85.15936	51.92006	72.86645	51.55132	64.02698	97.77291	62.65593	66.70309	70.79969	56.58228	55.01796	52.31301	33.63081	65.10367

CHPATER VI

Activity-Tolerated Blood Glucose Monitoring in Breath Sensing

6.1 Abstract

Breath sensing is a promising approach for non-invasively collecting various physiological data in real-time. However, the breath pattern is constantly disturbed by various daily activities. The concentration variation of the biomarker in exhaled breath is affected under the activities' influence.

Here we demonstrate an activity-tolerated blood glucose monitoring by artificial intelligence-based ensemble feature analysis of breath sensing data. In the researches, correlated multivariate signals, i.e., assembled features, were explored from the breath sensing data, correlated with blood glucose levels, and trained to predict blood glucose levels in the tested samples. By raising a thousand assembled features, the blood glucose levels were successfully expected even under the influences of various activities (e.g., alcohol drinking and exercising). The proof-of-concept was demonstrated in a breath print analysis using gas chromatography-mass spectrometry and extended to the study of breath sensing data collected using a chemiresistive sensor array. We successfully classified four levels of blood glucose level with 99.8 % accuracy using the breath sensing data collected under uncertain activities. Furthermore, we successfully demonstrated the blood glucose spikes monitoring under various activities, which had not been attainable by the conventional single biomarker-based method.

This activity-tolerated breath sensing opens a novel platform for investigating various human-to-environment interactions by collecting different physiological data under daily activities.

KEYWORDS: *Blood glucose, Breath analysis, GC-MS, Chemiresistive Sensor and Machine learning*

6.2 Introduction

Blood glucose (BG) monitoring contains the necessary information is glucose spikes, which can evaluate the signature of glucose metabolism. Furthermore, the glucose spike pattern allows us to predict the human activity and prognosis health conditions (i.e., fasting, consuming, training, stress, and others) because the coordinated physiological responses of multiple organs influence the variation of BG levels. Finally, glucose metabolism indicates insulin's competence and incompetence of glucose disposal into cells. Thus, daily-life glucose monitoring is a promising way to monitor health conditions, predict the symptoms and severe health problems, and prevent fatal disease (i.e., diabetes). In addition, monitoring can help people balance food or medicines intake related to a personal activity that leads to a healthy lifestyle. The most common method for BG monitoring is blood testing. Although BG checking by blood testing has high accuracy, this testing is inconvenient, painful, and at risk for infection, which is not accessible to everyone. To overcome this problem, recently, breath analysis has been developed, which has a great potential for non-invasive, real-time, and repetitive health diagnoses.

Breath analysis allows us to obtain the metabolic signature by monitoring and tracking the concentration variations of volatile biomarker molecules in exhaled breath. Many research works show the feasibility of BG prediction in fasting conditions by tracking specific biomarker molecules, such as acetone, ethanol, and xylene. However, these biomarkers also can produce by another metabolic pathway. It means that these biomarkers are easily fluctuated by various activities in daily life. For example, many research work proposed breath acetone as a high correlated biomarker of diabetic diagnostics—increasing breath acetone results from intense gluconeogenesis in the liver, which generates glucose from non-carbohydrate. Nevertheless, the liver also produces acetone during periods of caloric restriction of various scenarios such as fasting, intense exercise, and alcoholism. Furthermore, concentration variation of ethanol can be affected after drinking alcohol, and xylene can be affected by smoking. Thus, tracking specific biomarkers for BG monitoring/prediction might lead to a false prediction. Finally, we found that no report showed the achievement of practical BG prediction based on breath analysis/sensing during various activities before.

Here we demonstrated activity-tolerated blood glucose prediction in the breath analysis. We utilized the assembled biomarker combined with data preprocessing and machine learning to predict the multi-class BG levels under fasting, drinking, and exercise conditions. We further examined the real-time BG monitoring under fasting, drinking, and exercise with multi-channels of the chemo-sensitive resistor sensor array as a proof-of-concept with high potential, fast, and high stability for breath sensing.

6.3 Experimental Section

Breath Sample Collection and Blood Glucose Measurement. We collected the breath samples from the volunteers during oral glucose tolerance tests (OGTTs) under the three different activities, including i) fasting (for 8-10 h), ii) alcohol drinking (with Japanese distilled spirits), and iii) exercise (for 25 min). For the OGTTs, we utilized a 150 mL drink with 50 g glucose (TRELAN-G50, AY Pharmaceuticals) to control blood glucose. For the alcohol drinking and exercise activities, we employed 25 mL Japanese shochu 'Kuro-Kirishima (alcohol 25% contained) and a video training program of 'Focus T25 Speed 20 Beta', respectively, and the volunteers have experienced them just before taking the glucose drink. First, the exhaled breath was collected using a 10 L gas sample bag (Smart bag PA, GL Sciences). Then, for the breath component analysis by gas chromatography-mass spectrometry (GC–MS), we connected the gas sampling bag containing breath gas to an adsorbent-filled sample tube (Packed Liner with Tenax GR, mesh 80/100 #2414-1021, GL Science Inc.) and the breath gas of 500 mL was transferred to the sample tube using an air pump at a flow rate of 50 mL/min. The sample tubes were sealed and stored in the refrigerator at the temperature of -18 ºC until performing the GC–MS measurements. For the breath sensing using a chemiresistive sensor array, we directly connected the gas sampling bag with the sensing chamber. Then, the blood glucose value was measured together with collecting the exhaled breath during the OGTTs, by using a glucose meter through the conventional fingerstick method and a flash glucose monitoring system (FreeStyle Libre, Abbott).

Breath Component Analysis by GC-MS. We employed GC–MS (Shimadzu, GCMS-QP2020) equipped with an inlet temperature control unit (OPTIC) o carry out the collected breath gas component analysis. For the measurements, we used the inlet temperature at 300 ºC with split-less mode to thermally desorbed the collected breath gas in the sample tube. Then, the oven temperature was controlled to be 40 ºC for 5 min, elevated to 280 ºC with the rate of 5 ºC/min, and maintained at 280 ºC for 5 min. Then, we perform the InertCap 5MS/NP capillary column (60 m length, 0.25 mm inner diameter, 1 μm thickness, GL Science Inc.) to separate the desorbed compounds before transferring them to the MS system. We set the flow rates of helium gas (99.9999% pure) for column and purge to be 1 mL/min and 5 mL/min, respectively. We set the temperature of both the ion source and the interface of the mass-spectrometer to be 200 ºC during analysis. The mass to charge ratio (*m/z*) was characterized in the range of 35-300. Finally, we analyzed the obtained data by GCMS Solution ver. 4.45 SP1.

Fabrication of Chemiresistive Sensor Array. A chemiresistive sensor array was fabricated on a Si substrate by combining a lithographic patterning technique and a nano drop-casting method. Firstly, 16 pairs of comb-shaped Pt electrodes were patterned on a 7 × 7 mm^2 sized Si substrate (n-type, with 100 nm-thick SiO_2 cap layer) by photolithography and radio frequency (RF) sputtering. A 1 nm-thick Ti layer was used as an adhesive layer for Pt electrodes. The gap distance and the thickness of Pt electrodes were 40 μm and 400 nm, respectively. Then a 45 μm-thick SU-8 photoresist was coated on the electrode-patterned substrate by spin-coating and circular holes were made by photolithography. For the sensing materials, 16 types of GC stationary phase materials (GCSP) were chosen and the GCSP-carbon black (CB) nanocomposites were prepared by mixing 10 mg GCSP and 10 mg CB (45μm Graphitized carbon black, Sigma) in 10 mL *N*,*N*-dimethylformamide (DMF, Wako). The 16 types of GCSP were as follows: tetrahydroxyethylenediamine (THEED), GL Sciences/ *N*,*N*-Bis(2-cyanoethyl)formamide (BCEF), Tokyo Chemical Industry/LAC-3-R-728 (12% diethylene glycol succinate (DEGS), GL Sciences/DEGS, Supelco/ poly(ethylene succinate) (PES), Sigma/ UCON 75-H-90000, polyalkylene glycol (PAG) containing 75 wt% oxyethylene and 25 % oxypropylene groups, Shinwa Chemical Industries/1,2,3-Tris(2-cyanoethoxy)propane (TCEP), Supelco/SP-2330, poly (80% biscyanopropyl/20% cyanopropylphenyl siloxane), Supelco/SP-2340, poly (biscyanopropyl siloxane), Supelco/diglycerol, Tokyo Chemical Industry/Reoplex 400, GL Sciences/poly[di(ethylglycol)adipate] (PDEGA), Sigma/PEG4000, poly(ethyele glycol) 4000, Sigma/PEG20K, poly(ethyele glycol) 20000, United States Pharmacopeia (USP) Reference/PEG20M, poly(ethyele glycol) 20M, Shinwa Chemical Industries/free fatty acid phase (FFAP), Supelco. To form the GCSP-CB homogeneous suspension, the ultrasonic vibration at 38 kHz was applied for 60 min without adding any dispersant. The as-prepared GCSP-CB nanocomposite inks were drop-casted (40 nL total, 40 shots with the rate of 1 nL/shot) at the electrode patterned circular holes on the substrate by an ink-jet printing technique (custom-made, SIJ Technology Inc.). After depositing GCSP-CB nanocomposite inks, the device was sequentially annealed on a hotplate at 50 ºC for 60 min and in a vacuum oven (100 Pa) at 50 ºC for 60 min, and finally the 16-channel chemiresistive sensor array was obtained. The prepared devices were stored in a vacuum-sealed sample bag until conducting the breath sensing measurements. The microstructure of the sensor device was characterized by an optical microscopy (OLYMPUS DP21).

Breath Sensing Measurement by using 16-channel Chemiresistive Sensor Array. The breath sensing measurements were conducted by a homemade sensing system containing a sensing chamber, solenoid valves, an air pump and a sensor operation/data collection unit. For the measurements, the gas sampling bag containing exhaled breath gas was connected to the sensing chamber and the breath gas was transferred into the chamber by controlling the pumping rate at 100 mL/min of the air pump. The sensor responses were collected from 16 sensors as variations of output voltages when sequentially altering the flows of breath and N_2 carrier gas. The alternation of gas flows was performed every 10 s by controlling the solenoid valves using the sensor operation unit. The sensing response was defined as follows: $\Delta V/V_0=(V-V_0)/V_0$, where V_0 and V are the output voltages under the flows of N_2 carrier gas and breath gas, respectively. All breath sensing measurements were performed at room temperature.

Data Analysis of Breath Sensing by Artificial Intelligence. Before the data analysis, the baseline correction was performed for the obtained sensing curves. The sensing responses ΔV were collected from a 16-channel sensor array and used as the dataset for machine learning. One thousand sixty-eight datasets of sensing response were obtained for each person (3,204 sensing data for three people). For machine learning, a random forest algorithm was employed to build classifiers. The hyper-parameters optimized the multi-layer perceptron (MLP)models. A 5-fold cross-validation was used to confirm the reproducibility of a classifier. The average area characterized the classifier's reliability under the curve (AUC) of the receiver operating characteristic (ROC) curve. Finally, the prediction accuracy and the coefficient of variation in prediction accuracy were computed to evaluate the performance of breath odor sensing-based individual authentication.

Data Analysis of Breath Component by Artificial Intelligence. A 2-dimensional (2D) molecular profile as retention time and *m/z* was created by merging the mass chromatograms with log-scaled intensity (256-levels) in the range of *m/z* 35-300. Intensities of the molecular fragment peaks in the 2D molecular profile were extracted together with their peak positions by data mining and image processing. The datasets for each molecular fragment peak were made and evaluated with the blood glucose level by machine learning using an artificial neural network (ANN) with Keras/Tensor flow application programming interface (API). Prediction accuracy of blood glucose level was calculated by increasing the number of used molecular fragment peaks to create the predictive model. We rearranged the list of molecular fragment peaks based on p-value, retention time, and peak intensity for these calculations and evaluated rearrangement's impact on the prediction accuracy. The p-value for each fragment peak was obtained from paired sample t-test of the lowest and the highest blood glucose conditions data. Finally, we evaluated the performance of created predictive models via the prediction accuracy, sensitivity, the performance of created predictive models were evaluated by prediction accuracy, sensitivity, specificity, confusion matrix, and area under the curve (AUC) of receiver operating characteristics (ROC) curve, which were calculated and optimized by Python ver.3.7.7.

6.4 Results and Discussion

We analyzed the chemical components of the collected samples using gas chromatography linked with mass spectrometry (GC-MS) (Figure 1a). Two-dimensional GC-MS (2D GC-MS) images based on GC-MS provide the assembled biomarker information for BG prediction and identification of the biomarker pattern.

Two-dimensional GC-MS (2D GC-MS) image shows a molecular fragment pattern between low (≤120 mg/dL) and high (>120 mg/dL) BG levels in fasting, drinking, and exercise conditions (Figure 1b). Based on 2D GCMS image and image processing, ~7,980 molecular fragment peaks were detected. This molecular fragment peak pattern was related to a human metabolic pathway. For the fasting condition, the clear difference between high and low BG levels was seen in these peak patterns and shown in a 2D differential image for each condition. The 2D differential image allows us to easily observe the increasing and decreasing of molecular fragment peaks when BG levels increase. We found that BG metabolism in the case of BG increases the effect on a changing of many biomarkers. Moreover, fasting, drinking, and exercise conditions showed differences in molecular fragment peak patterns after BG levels increased.

Since detected molecular fragment peaks contain both molecular fragment peaks, which are significant and non-significant with BG levels, before applying the assembled biomarker, data set into machine learning, molecular fragment peak screening was required. Therefore, this study applied the p-value to screen the non-significant molecular fragment peak out. The molecular fragment peaks were rearranged according to ascending p-value from the paired sample t-test between low and high BG levels (Figure 2a). Based on our algorithm, we can detect the molecular fragment peak, which has a p-value lower than 0.05 for fasting condition, mixed conditions of fasting/drinking and

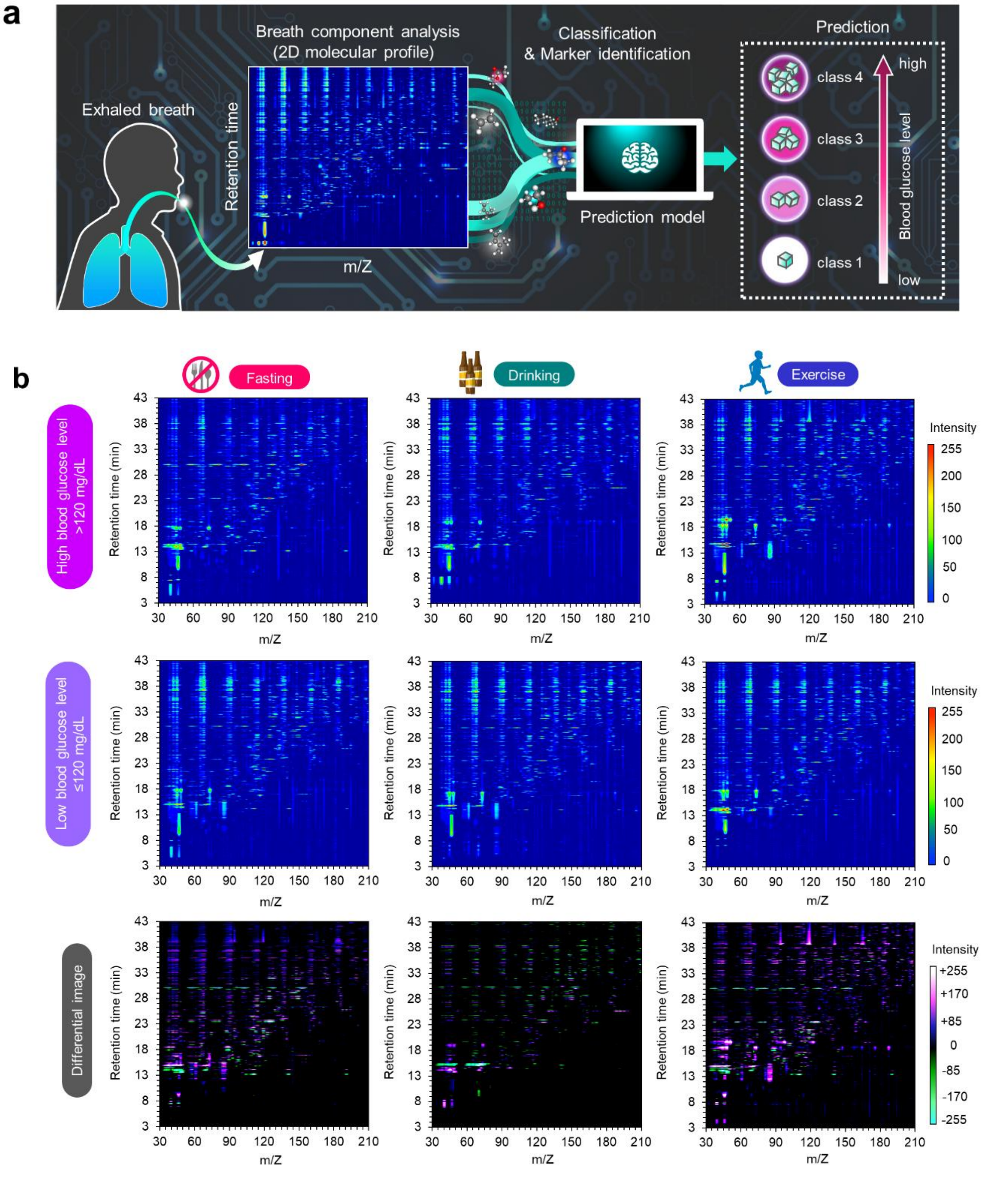


Figure 1. (a) Schematic illustration of breath component analysis and blood glucose level prediction using artificial intelligence approach. (b) 2D molecular profiles in exhaled breath obtained at high (>120 mg/dL) and low (≤120 mg/dL) blood glucose levels under

various activities (i.e. fasting, alcohol drinking, exercise). The differential images of 2D molecular profiles at high and low blood glucose levels are shown in the lower row.

fasting/drinking/exercising were 790, 959, and 1,270 peaks, respectively (standard threshold should be set at p<0.05). It means that we can detect a large number of significant biomarkers which correlate with BG levels. These results clarify that the assembled biomarker can overcome the low robustness of BG prediction as investigated from previous research.

To investigate the performance of the assembled biomarker, we applied machine learning to the data set (peak position and peak intensity) and evaluated the prediction accuracies of BG levels in two-class. When increasing the utilized number of molecular fragment peaks, the prediction accuracies for fasting condition, mixed conditions of fasting/drinking and fasting/drinking/exercising trend to be increasing reached to 100.0%, 100.0%, 97.9% by using 28, 371, and 892, respectively (Figure 2b, and Table I). However, the prediction accuracy trend decreases after using over 1,500 peaks for BG prediction. This result related to the molecular fragment peak, which was ordered by p-value that the low significant and non-significant molecular fragment peaks were contained in the data set after ~1,300 peaks. These molecular fragment peaks lead to fatal error in the predictive model, decreasing prediction accuracy. However, we can archive to discriminate BG levels in two-class by increasing the utilized number of fragment peaks although the complexity of the condition was increased. These results showed that utilizing the number of molecular fragment peaks and the assembled biomarker plays an essential role for robust and effective BG prediction. The performance of predictive models for fasting, fasting/drinking, and fasting/drinking/exercising conditions were shown in the area under the curve (AUC) of the receiver operating curve (ROC) (Figure 1d).

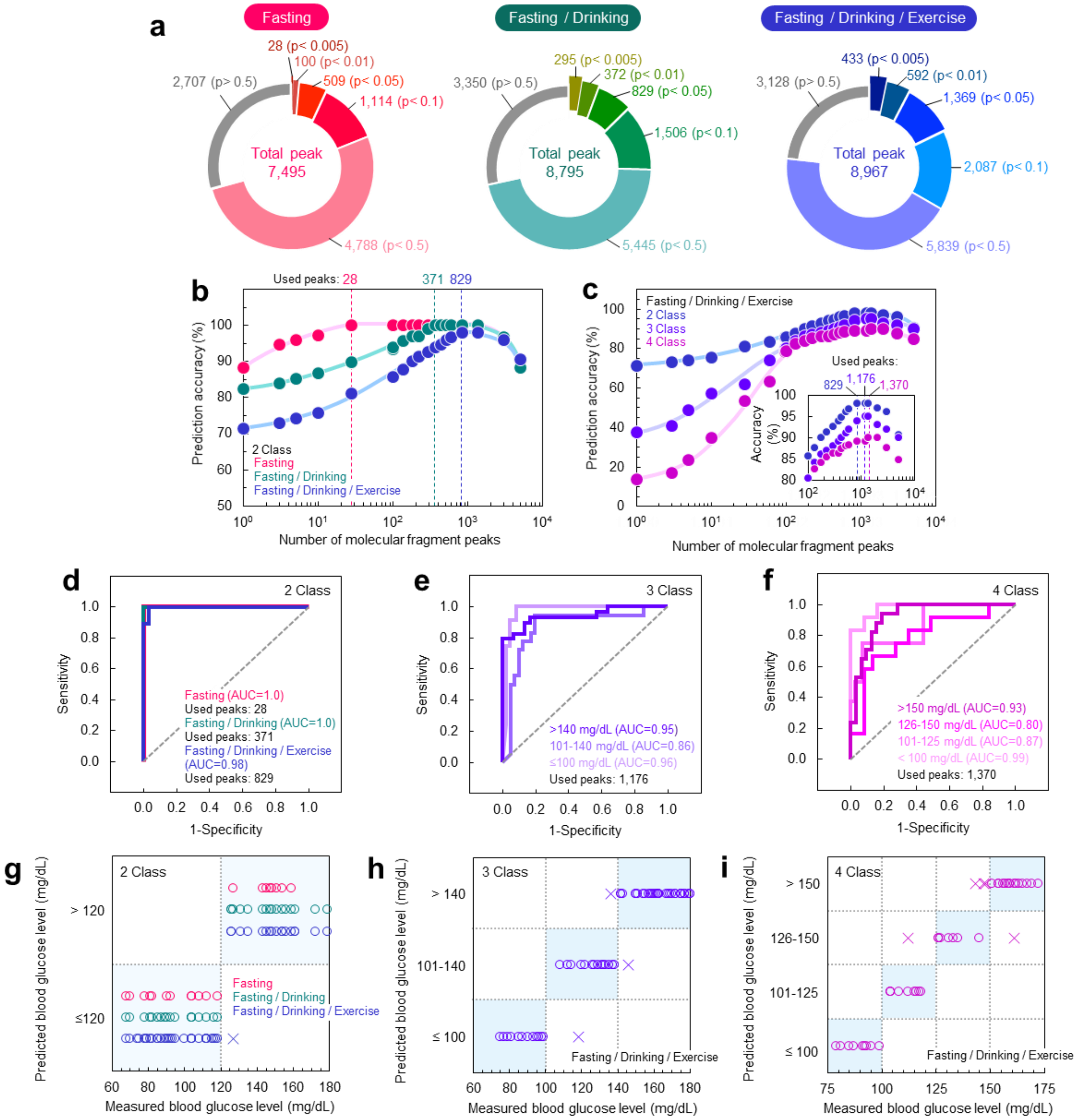


Figure 2. (a) Number of molecular fragment peaks in 2D molecular profiles of exhaled breath categorized by p-value (left: fasting, middle: fasting/drinking, right: fasting/drinking/exercise). (b) Prediction accuracies of 2 class blood glucose level discrimination (>120 mg/dL, ≤120 mg/dL) as a function of number of used molecular

fragment peaks in the conditions of fasting (pink), fasting/drinking (green) and fasting/drinking/exercise (blue), respectively. The minimum number of used peaks for achieving the highest prediction accuracy in each condition are noted. (c) Prediction accuracies of 2 class (blue), 3 class (violet) and 4 class (purple) blood glucose level discriminations (2 class: ≤120 mg/dL, >120 mg/dL, 3 class: ≤100 mg/dL, 101-140 mg/dL, >140 mg/dL, 4 class: ≤100 mg/dL, 101-125 mg/dL, 126-150 mg/dL, >150 mg/dL) as a function of number of used molecular fragment peaks in the mixed conditions of fasting/drinking/exercise. Inset shows the magnified data around the highest prediction accuracy. The minimum number of used peaks for achieving the highest prediction accuracy in each condition are noted. (d-f) ROC curves of classifiers used for discriminating the blood glucose levels: (d) 2 class, (e) 3 class and (f) 4 class. The number of used peaks for creating the classifier and the AUC values for each classifier are noted. (g-i) Confusion matrixes of exhaled breath based blood glucose level prediction represented by symbols (○: true prediction, ×: false prediction): (g) 2 class, (h) 3 class and (i) 4 class.

Table I. Data summary of 2 class blood glucose level discrimination via breath component analysis in various conditions (i.e. fasting, fasting/drinking, fasting/drinking/exercise).

Conditions	Used peaks	Training data	Validation data	Testing data	Accuracy
fasting	28	26	6	20	100 %
fasting/drinking	371	49	15	35	100 %
fasting/drinking/exercise	829	148	36	49	97.9 %

Based on our calculation, we challenged the multi-classification applied to this study. In this case, the three-class and four-class of mixed condition fasting/drinking/ exercising were predicted (Figure 2c, 2e, 2f, and Table II). The predictive result showed that we predicted four-class BG predictions with high accuracy. However, we found that obtaining the optimized condition using 829 peaks of the molecular fragment was not enough. Therefore, the utilized molecular fragment peaks increased to 1,176 and 1,370, respectively, to increase the accuracy of three-class and four-class predictions. It proves that increasing the utilized number of molecular fragment peaks can predict the complicated condition. The result clearly showed that increasing an assembled biomarker plays a vital role in robust BG prediction and discriminates the different classes of BG levels.

Table II. Data summary of 2 class, 3 class and 4 class blood glucose level discriminations via breath component analysis in the mixed conditions of fasting/drinking/exercise.

Conditions	Used peaks	Training data	Validation data	Testing data	Accuracy
2 class	829	148	36	49	97.9 %
3 class	1,167	148	36	49	95.9 %
4 class	1,370	148	36	49	89.8 %

.

In order to confirm the robustness and efficiency of the prediction result, we identified the chemical composition of the top 20 biomarkers in each condition. It evaluated the individual prediction accuracy for two-class BG prediction by machine learning. The chemical component analysis revealed that the top 20 biomarkers in fasting condition (ordered by p-value) was disturbed and rearranged ~90% in case of fasting/drinking and

~85% in case of fasting/drinking/exercising conditions due to the interference by the metabolism of drinking and exercise (Figure 3a, 3b, and 3c). The result indicated that the biomarker pattern was changed depending on human metabolism. For evaluating the individual performance of each biomarker, the average prediction accuracy of the top 20 biomarkers for the fasting condition was 77.6%, mixed condition of fasting/drinking was 76.2%, and fasting/drinking/exercise was 69.3%. The individual prediction accuracy for each biomarker of each condition shows a trend to be decreased when the complexity of the condition increases (Figure 3a, 3b, and 3c). This analysis found that alkene, aldehyde, alcohol, ketone, and aromatic compounds show the most significance for blood glucose prediction. Furthermore, considering the individual accuracy of each biomarker combined with the identification of a significant functional group, we found that the single biomarker cannot obtain high prediction accuracy in the complicated condition. In contrast, the assembled biomarker with various functional group species increases the prediction accuracy, resulting in robust and effective BG prediction. Strong evidence that focuses on specific biomarkers as investigated in the previous study leads to an error for BG prediction in a complicated condition.

In other words, a highly accurate and robust BG analysis/sensing even under various life activities would be possible by monitoring the assembled-biomarker molecules. For proving this concept, we performed the BG monitoring via multi-channels of the chemo-sensitive resistor sensor array (Figure 4a), of which each sensor can detect a variety of molecules. The sixteen sensors were sensing the three conditions, including fasting, drinking, and exercise, with four different BG levels. The change in voltage was measured in $\Delta V/V0$, where $\Delta V/V0$ is the change in voltage from baseline to the highest voltage response divided by the baseline voltage. Each sensor showed a quick response/recovery and high reproducibility for breath sensing (Figure 4a).

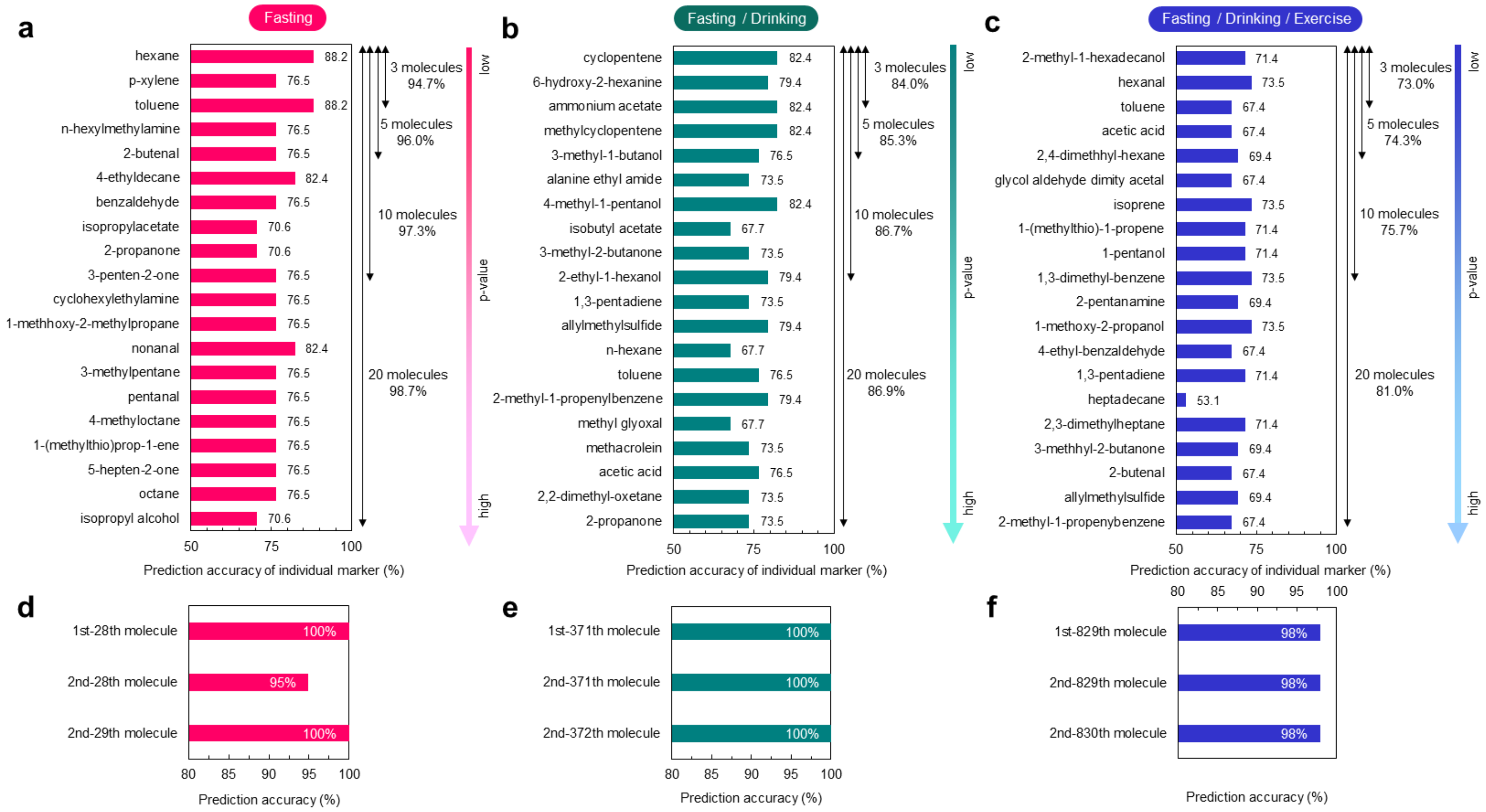

a
Fasting
hexane 88.2
p-xylene 76.5
toluene 88.2
n-hexylmethylamine 76.5
2-butenal 76.5
4-ethyldecane 82.4
benzaldehyde 76.5
isopropylacetate 70.6
2-propanone 70.6
3-penten-2-one 76.5
cyclohexylethylamine 76.5
1-methhoxy-2-methylpropane 76.5
nonanal 82.4
3-methylpentane 76.5
pentanal 76.5
4-methyloctane 76.5
1-(methylthio)prop-1-ene 76.5
5-hepten-2-one 76.5
octane 76.5
isopropyl alcohol 70.6
50
75
100
Prediction accuracy of individual marker (%)
3 molecules 94.7%
5 molecules 96.0%
10 molecules 97.3%
20 molecules 98.7%
low
p-value
high
b
Fasting / Drinking
cyclopentene 82.4
6-hydroxy-2-hexanine 79.4
ammonium acetate 82.4
methylcyclopentene 82.4
3-methyl-1-butanol 76.5
alanine ethyl amide 73.5
4-methyl-1-pentanol 82.4
isobutyl acetate 67.7
3-methyl-2-butanone 73.5
2-ethyl-1-hexanol 79.4
1,3-pentadiene 73.5
allylmethylsulfide 79.4
n-hexane 67.7
toluene 76.5
2-methyl-1-propenylbenzene 79.4
methyl glyoxal 67.7
methacrolein 73.5
acetic acid 76.5
2,2-dimethyl-oxetane 73.5
2-propanone 73.5
50
75
100
Prediction accuracy of individual marker (%)
3 molecules 84.0%
5 molecules 85.3%
10 molecules 86.7%
20 molecules 86.9%
low
p-value
high
c
Fasting / Drinking / Exercise
2-methyl-1-hexadecanol 71.4
hexanal 73.5
toluene 67.4
acetic acid 67.4
2,4-dimethhyl-hexane 69.4
glycol aldehyde dimity acetal 67.4
isoprene 73.5
1-(methylthio)-1-propene 71.4
1-pentanol 71.4
1,3-dimethyl-benzene 73.5
2-pentanamine 69.4
1-methoxy-2-propanol 73.5
4-ethyl-benzaldehyde 67.4
1,3-pentadiene 71.4
heptadecane 53.1
2,3-dimethylheptane 71.4
3-methhyl-2-butanone 69.4
2-butenal 67.4
allylmethylsulfide 69.4
2-methyl-1-propenybenzene 67.4
50
75
100
Prediction accuracy of individual marker (%)
3 molecules 73.0%
5 molecules 74.3%
10 molecules 75.7%
20 molecules 81.0%
low
p-value
high
d
1st-28th molecule 100%
2nd-28th molecule 95%
2nd-29th molecule 100%
80
85
90
95
100
Prediction accuracy (%)
e
1st-371th molecule 100%
2nd-371th molecule 100%
2nd-372th molecule 100%
80
85
90
95
100
Prediction accuracy (%)
f
80
85
90
95
100
1st-829th molecule 98%
2nd-829th molecule 98%
2nd-830th molecule 98%
Prediction accuracy (%)

Figure 3. (a-c) List of 1st-20th biomarker molecules in exhaled breath used for 2 class blood glucose level discrimination in the conditions of (a) fasting, (b) fasting/drinking and (c) fasting/drinking/exercise, respectively. The molecules are arranged by the p-value. The prediction accuracies of individual molecule, 1st-3rd molecules, 1st-5th molecules, 1st-10th molecules and 1st-20th molecules are noted. (d-f) Prediction accuracies of 2 class blood glucose level discrimination when intentionally excluding the specific biomarkers for the analysis. The conditions of (d) fasting, (e) fasting/drinking and (f) fasting/drinking/exercise.

The average sensor response showed each condition's different sensor patterns for each BG level (Figure 4b). To show the concept which consistency with the GC-MS result, the number of utilized sensors were increased result in increasing of prediction accuracies for fasting condition, mixed conditions of fasting/drinking and fasting/drinking /exercising reached 98.1%, 96.0%, and 92.9%, respectively (Figure 5c, 5d, 5e, 5f, and Table III). However, comparing the prediction accuracies from GC-MS with multi-channels of the chemo-sensitive resistor sensor array is not consistent because while GC-MS required over 1,370 features to achieve the best prediction (89.8% accuracy) for fasting/drinking /exercise condition but multi-channels of the chemo-sensitive resistor sensor array used only 16 features to achieve the best prediction (92.9% accuracy).

Thus, which factor is a cause of this phenomenon? To clarify this phenomenon, we tried to decrease the number of training data for the predictive model due to the number of training data for multi-channels of the chemo-sensitive resistor sensor array higher than GC-MS ~21 folds which might be the main effect of this phenomenon. By decreasing training data of multi-channels of the chemo-sensitive resistor sensor array to ~100 data (same as GC-MS predictive model), the prediction accuracies decrease to 55.0%. The result proved that higher accuracy of multi-channels of the chemo-sensitive resistor sensor array is affected by training data and confirms the consistency of prediction result between GC-MS and the chemo-sensitive resistor sensor array.

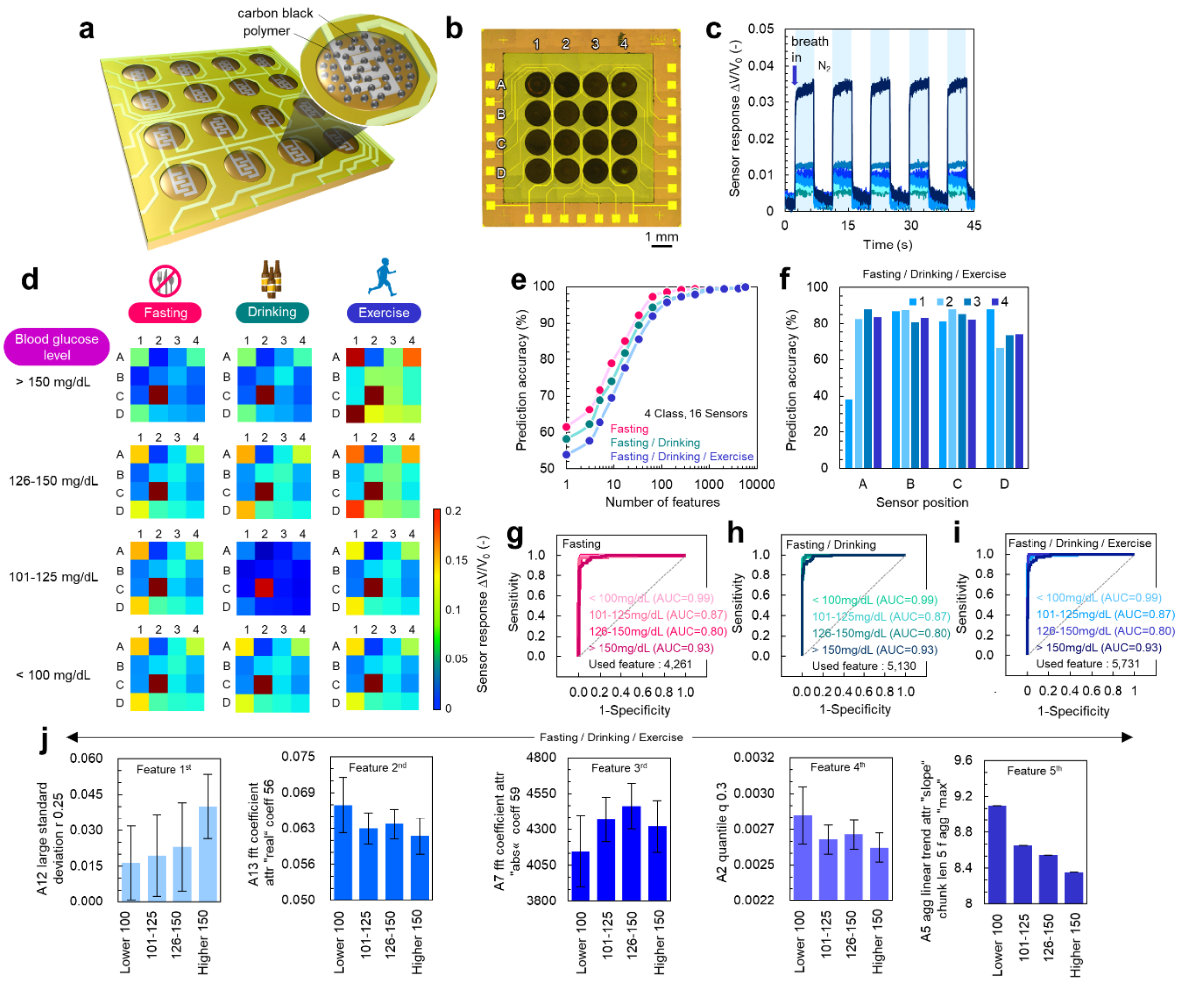


Figure 4. (a) Schematic illustration and (b) optical microscopy image of 16 channel chemiresistive sensor array. The alphabets (A-D) and the numbers (1-4) represent the names of rows and columns for addressing each sensor. (c) Typical five-successive sensor

responses $\Delta V/V_0$ of each sensor when introducing the exhaled breath sample. The sensing measurement was performed at room temperature by pumping breath sample and N_2 gas sequentially. (d) Average sensor responses $\Delta V/V_0$ of each sensor for 4 class blood glucose level discrimination in the conditions of fasting, drinking, exercise, respectively. (e) Prediction accuracies of 4 class blood glucose discrimination as a function of number of used sensors in the conditions of fasting (pink), fasting/drinking (green) and fasting/drinking/exercise (blue), respectively. (f) Prediction accuracies of individual sensor. (g-i) ROC curves of classifiers used for discriminating the blood glucose levels in the conditions of (g) fasting, (h) fasting/drinking and (i) fasting/drinking/exercise, respectively. The AUC values for each classifier are noted.

Table III. Data summary of 4 class blood glucose level discrimination via breath gas sensing in various conditions (i.e. fasting, fasting/drinking, fasting/drinking/exercise).

Conditions	Training data	Validation data	Testing data	Accuracy
fasting	410	280	376	98.1 %
fasting/drinking	819	560	752	96.0 %
fasting/drinking/exercise	1,229	840	1,127	92.9 %

Finally, our approach was applied to real-time BG monitoring, including fasting, drinking, and exercise (Figure 5a, 5b, and 5c). We achieved to monitor BG levels with high accuracy. Besides, the difference in glucose spike patterns was recognized. In the case of fasting conditions, we can see healthy people's simple glucose spike pattern related to blood testing (Figure 5a). The exciting thing is that the delay of a glucose spike under drinking conditions (Figure 5b) and reduction of a glucose spike under exercise conditions (Figure

5c) were observed. The different glucose spike was caused by drinking and exercise, reported in the previous study. Moderate alcohol consumption decreases insulin which affects decreasing glucose metabolism. At the same time, exercise can improve insulin action, increasing glucose metabolism.

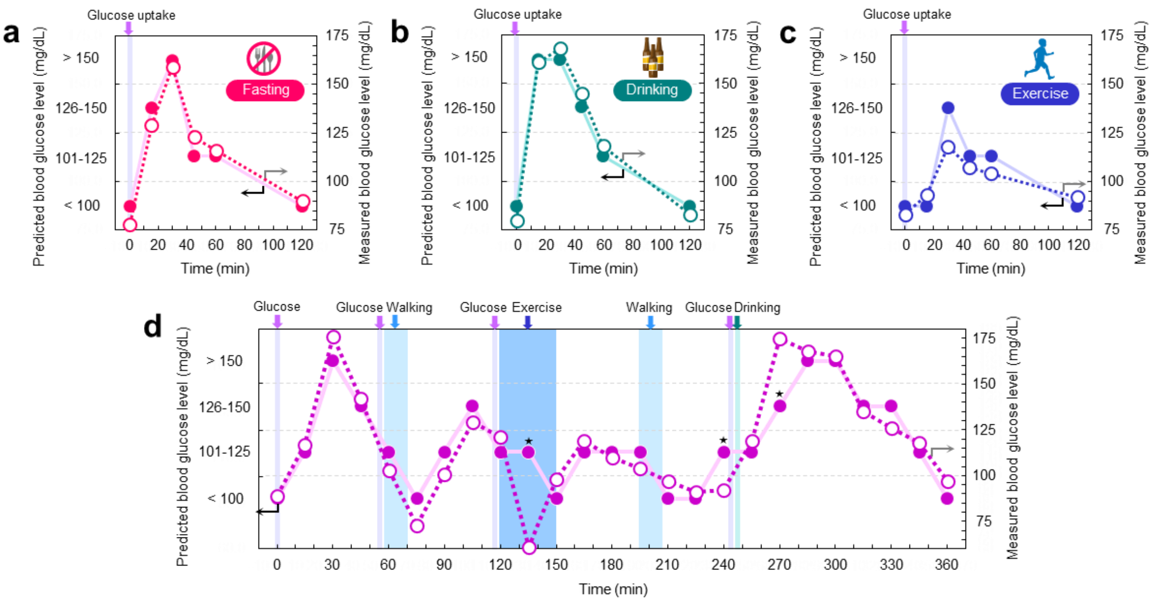


Figure 5. (a-c) Profiles of the predicted blood glucose levels via breath gas sensing and the measured blood glucose levels via glucose sensor after taking glucose in the conditions of (a) fasting, (b) alcohol drinking and (c) exercise, respectively. (d) Profiles of the predicted blood glucose levels via breath gas sensing and the measured blood glucose levels via glucose sensor under the dynamically varying activities (i.e. glucose uptake, walking, exercise, drinking). For these experiments, the classifier of 4 class blood glucose level discrimination was utilized.

6.5 Conclusions

In conclusion, this study provides a robust and effective system for BG prediction based on GC-MS and multi-channels of the chemo-sensitive resistor sensor. GC-MS analysis of exhaled breath revealed that the biomarker pattern of BG was disturbed depending on human metabolism. To figure out this problem by utilizing the assembled biomarker, we can predict multi-class BG prediction with high accuracy under complicated conditions. Moreover, we succeeded in real-time BG monitoring under various conditions, including fasting, drinking, and exercising. The different glucose spike patterns were seen under fasting, drinking, and exercising conditions, consistent with a previous study of blood tests based on glucose monitoring. Thus, these preliminary results provide a concept of robust breath analysis/monitoring via the molecular sensor array towards non-invasive collecting the chemical information in the human body and understanding the complex metabolic interaction with the environment.

6.6 References

1. Röder PV, Wu B, Liu Y, Han W. Exp Mol Med. 2016;48(3): e219.
2. Blaak EE, Antoine JM, Benton D, et al. Obes Rev. 2012;13(10):923-984.
3. Rghioui, A.; Lloret, J.; Harane, M.; Oumnad, A. *Electronics* 2020*, 9*, 678.
4. Heinemann L, Stuhr A, Brown A, et al. Eur Endocrinol. 2018;14(2):24-29.
5. Cappon, G.; Acciaroli, G.; Vettoretti, M.; Facchinetti, A.; Sparacino, G. *Electronics* 2017*, 6*, 65.
6. Roberts K, Jaffe A, Verge C, Thomas PS. J Diabetes Sci Technol. 2012;6(3):659-664.
7. Yan K, Zhang D. Annu Int Conf IEEE Eng Med Biol Soc. 2014; 2014: 6406-9.
8. Bodini A, D'Orazio C, Peroni D, Corradi M, Folesani G, Baraldi E, Assael BM, Boner A, Piacentini GL. Pediatr Pulmonol. 2005;40(6):494–499.
9. Pham, Y.L.; Beauchamp, J. *Molecules* 2021*, 26*, 5514.
10. Turner C, Walton C, Hoashi S, Evans M. J Breath Res. 2009 Dec;3(4):046004.
11. Rydosz A. J Diabetes Sci Technol. 2015;9(4):881-884.
12. Galassetti PR, Novak B, Nemet D, Rose-Gottron C, Cooper DM, Meinardi S, Newcomb R, Zaldivar F, Blake DR. Diabetes Technol Ther. 2005 Feb;7(1):115-23.
13. Refz, P., Obermeier, J., Lehbrink, R. et al. Sci Rep 9, 15707 (2019).
14. Leopold JH, van Hooijdonk RT, Sterk PJ, Abu-Hanna A, Schultz MJ, Bos LD. BMC Anesthesiol. 2014; 14:46.
15. Leopold JH, Bos LDJ, Colombo C, Sterk PJ, Schultz MJ, Abu-Hanna. J Breath Res. 2017 Apr 7;11(2):026002

CHAPTER VII

OVERALL CONCLUSION AND OUTLOOK

Big data in chemistry allow us to discover new chemical information in diverse field. The concept of Big data in chemistry is a massive collection of data (Volume), high-speed processing and fast data collection (Velocity), and diverse characteristic of data (Variety). The analysis and interpretation following this concept make Big data valuable, particularly for the new era of chemistry. Odor sensing can be a candidate promising to gain and accumulate the chemical information's vast and varied data set, which satisfies the concept of Big data. Therefore, analyzing enormous chemical information in odor sensing is critical for Big data in chemistry. This thesis focuses on developing the analytical chemistry for chemical analysis/sensing and accomplishing breakthroughs in scientific and technological areas.

Chapter III presented a method named NPFimg, which automatically identifies the multivariate chemo-/biomarkers feature of analytes in chromatography−MS data without the peak picking process, a critical problem for raw MS data processing. NPFimg utilizes the synergetic between image processing and machine learning and processes a 2D MS map to discriminate analytes and identify and visualize marker features. NPFimg allows us to comprehensively characterize the signals in MS data without employing the conventional peak picking process, which suffers from false peak detections. We successfully demonstrated the feasibility of chemo-/biomarker characterization in case studies of aroma odor and human breath on GC−MS, even at the ppb level. Comparison with the widely used automated peak detection algorithm such as XCMS showed the excellent reliability of NPFimg, in that it had lower error rates of the signal acquisition and the feature identification of chemo-/ biomarkers. In addition, we showed the potential applicability of NPFimg to the untargeted metabolomics of human breath. NPFimg is potentially applicable to diverse metabolomics/chemometrics data processing GC− and LC−MS. Because the time cost in NPFimg is much shorter than the peak picking-based conventional approaches. The high throughput online MS data analysis of various complex analytes would be expected by uploading the data file on Cloud space.

Chapter IV demonstrated the preliminary study of breath odor sensing-based individual authentication using an artificial olfactory sensor array. We used a 16-channel chemiresistive sensor array to test the breath odor samples collected from 6 people and then analyzed the acquired sensing responses by machine learning with a random forest algorithm. We successfully achieved the individual authentication of 6 persons with the median accuracy of 96.4 %. In addition, we found that the accuracy and the reproducibility significantly improved by increasing the number of used sensors. We demonstrated the breath odor sensing-based individual authentication for the fasted subjects in this study. However, it remains a challenging issue to demonstrate its feasibility under the interferences of disease-related metabolites and exogenous compounds originating from the diets and the tested environment the practical application. The barrier must be overcome by utilizing multiple sensors and extracting numerous features from the sensing curves. Nevertheless, we believe that our findings in this study provide an essential foundation for breath odor sensing-based biometrics.

Chapter V demonstrates an activity-tolerated blood glucose monitoring by artificial intelligence-based ensemble feature analysis of breath sensing data. In the analyses, correlated multivariate signals, *i.e.*, ensemble features, were explored from the breath sensing data, correlated with blood glucose levels, and trained to predict blood glucose levels in the tested samples. By introducing a thousand ensemble features, the blood glucose levels were successfully predicted even under the influences of various activities (e.g., alcohol drinking and exercising). The proof-of-concept was demonstrated in a breath print analysis using gas chromatography-mass spectrometry and extended to the analysis of breath sensing data collected using a chemiresistive sensor array. We achieved a four-level classification of blood glucose level with 99.8 % accuracy using the breath sensing data collected under uncertain activities. Furthermore, we successfully demonstrated the blood glucose spikes monitoring under ambulating conditions, which had not been attainable by the conventional single biomarker-based method. This activity-tolerated breath sensing opens a novel platform for investigating various human-to-environment interactions by collecting various physiological data under daily activities.

Although many researchers are devoted to realizing the actual application of odor sensing to obtain the concept of Big data in chemistry, it still has many challenging issues for breath odor analysis development because the development of breath analysis requires a multiplicity of technology—for example, analytical chemistry, biology, material science, and sensor engineering. Our study shows the impact of various features on activity-tolerated physiological data in a complex mixture like human breath. Furthermore, it means that a diverse character of the sensor is necessary for sensing a wide range of chemical information of odor with high quality and low error rate. Therefore, the development of new feature extraction and feature analysis methods is an excellent choice to analyze odor sense if a variety of sensors is limited.

LIST OF PUBLICATIONS

LIST OF PUBLICATIONS

Scientific Journals

【***Scientific Papers-For Ph.D.***】

1. Image Processing and Machine Learning for Automated Identification of Chemo-/Biomarkers in Chromatography–Mass Spectrometry

   C. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, W. Tanaka, B. Samransuksamer, G. Zhang, J. Liu, M. Kanai, and T. Yanagida

   Analytical Chemistry 2021, 93, 14708−14715

   Selected as Supplementary Cover

2. Discrimination of Complex Odors with Gas Chromatography-Mass Spectrometry Data by Texture Image Analysis and Machine Learn

   C. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, W. Tanaka, M. Kanai, and T. Yanagida

   Engineering Sciences Reports 2022, 43, 42-48

3. Breath Odor Based Individual Authentication by Artificial Olfactory Sensor Array and Machine Learning

   C. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, B. Samransuksamer, Y. Hanai, A. Nakao, M. Nakatani, J. Liu, G. Zhang, W. Tanaka, M. Kanai, T. Yasui, Y. Baba, and T. Yanagida

   in submission

4. Robust Collection and Monitoring of Physiological Data by Ensemble Feature Analysis of Breath Sensing Data

   C. Jirayupat, K. Nagashima, T. Takahashi, T. Hosomi, J. Liu, G. Zhang, W. Tanaka, B. Samransuksamer, M. Kanai, A. Nakao, Y. Hanai, M. Nakatani, T. Yasui, Y. Baba and T. Yanagida

   in preparation

【***Scientific Papers-For References***】

5. Identification of Genetic Variants via Bacterial Respiration Gas Analysis

   N. Koga, T. Hosomi, M. Zwama, **C. Jirayupat**, T. Yanagida, K. Nishino and S. Yamasaki

   Frontiers in Microbiology 2020, 11, 581571.

Conferences

1. Development of Data Mining Methodology for Breath Analysis (Poster)

   C. Jirayupat, K. Nagashima, T. Takahashi, G. Zhang, M. Kanai and T. Yanagida

   International Conference on Science and Technology of Emerging Materials 2018 (STEMa2018), Pattaya, Thailand, 2018 July 18-20.

2. アンサンブル分子センシングによる堅牢な呼気診断法の開発 (口頭発表)

   長島 一樹, **C. Jirayupat**, 細見 拓郎, 高橋 綱己, G. Zhang, 金井 真樹, 柳田 剛

   第81回応用物理学会秋季学術講演会, オンライン, 2020年9月8-11日

Patent

1. 出願番号：特願2018-168625

   発明者：柳田　剛, 長島　一樹, 高橋　綱己, 細見　拓郎, **C. Jirayupat**

   発明の名称：「推定装置及び推定方法」

   出願人：九州大学

   出願日：2018年9月10日

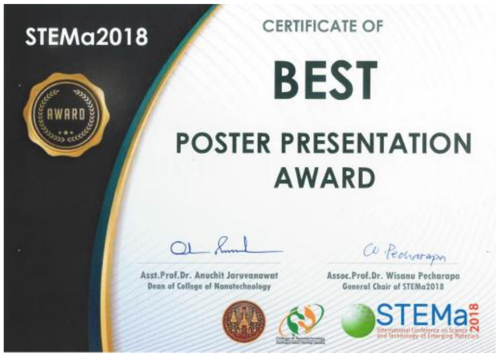
Award

1. Best Poster Presentation Award

   C. Jirayupat

   "Development of Data Mining Methodology for Breath Analysis"

   International Conference on Science and Technology of Emerging Materials 2018 (STEMa2018), Pattaya, Thailand, 2018.7.20



STEMa2018

AWARD

CERTIFICATE OF

BEST

POSTER PRESENTATION AWARD

Asst.Prof.Dr. Anuchit Jaruvanawat
Dean of College of Nanotechnology

W. Pecharapa
Assoc.Prof.Dr. Wisanu Pecharapa
General Chair of STEMa2018

STEMa 2018
International Conference on Science and Technology of Emerging Materials