analytical chemistry

pubs.acs.org/ac Article

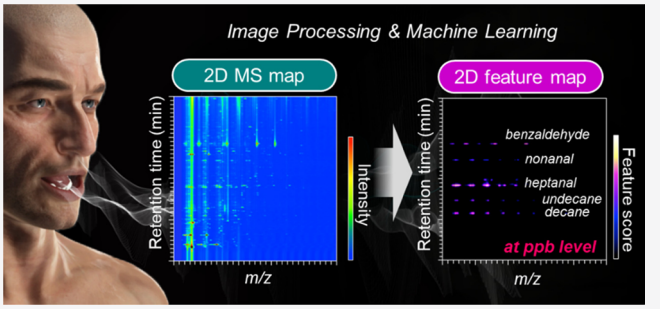
Image Processing and Machine Learning for Automated Identification of Chemo-/Biomarkers in Chromatography−Mass Spectrometry

Chaiyanut Jirayupat, Kazuki Nagashima,* Takuro Hosomi, Tsunaki Takahashi, Wataru Tanaka, Benjarong Samransuksamer, Guozhu Zhang, Jiangyang Liu, Masaki Kanai, and Takeshi Yanagida*

Cite This: https://doi.org/10.1021/acs.analchem.1c03163 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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.



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 above-mentioned 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

Received: July 26, 2021
Accepted: October 18, 2021

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

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 a 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.

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 with the vaporized aroma odors. The vial bottle has two 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 adsorbent-filled 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.

GC−MS MEASUREMENT

Mass chromatogram data of the aroma odor samples and the human breath samples were obtained by GC−MS (GCMS-QP2020, 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

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX

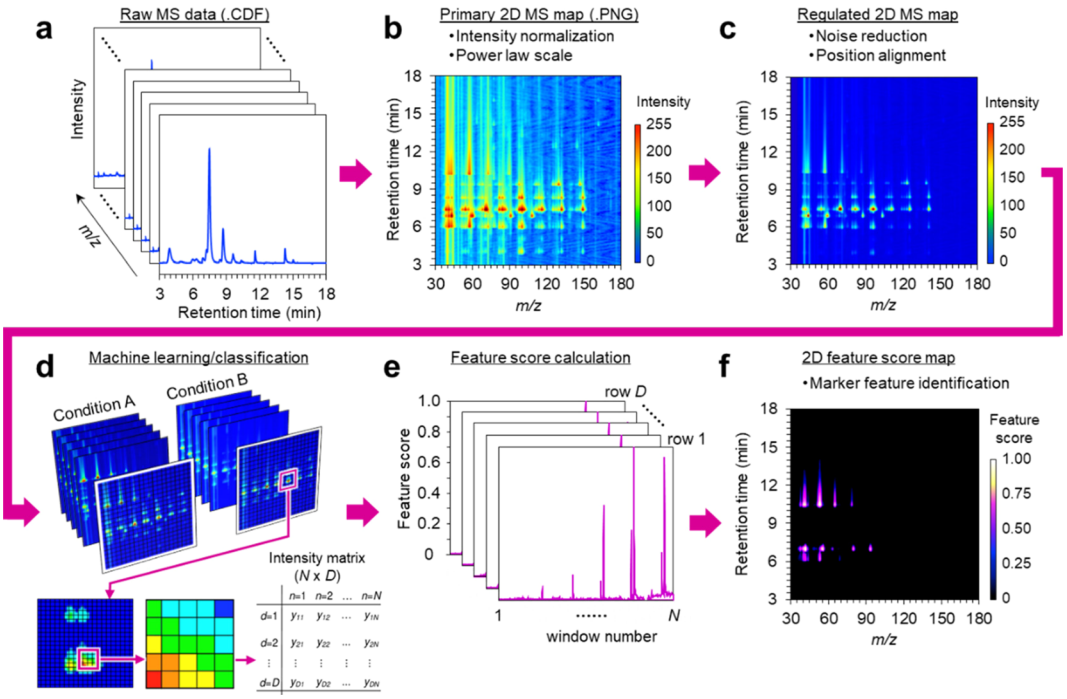


Figure 1. Graphical workflow of *NPFimg* for visualizing chemo-/biomarker signals from raw gas chromatography−mass spectrometry (GC−MS) data. Starting from (a) series of raw MS data, the workflow follows (b) creation of a primary two-dimensional (2D) MS map (m/z vs retention time) with power-law scale 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.

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.

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 (x-axis) 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 (γ = 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: $\log p/(1-p) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \ldots + x_n\beta_n$, where p is the probability of which the data sets can be classified, x_n is the intensity of each segment in the 2D MS map, β_n ($n \geq 1$) is the model's learned weight (i.e., feature score), and β_0 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

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX

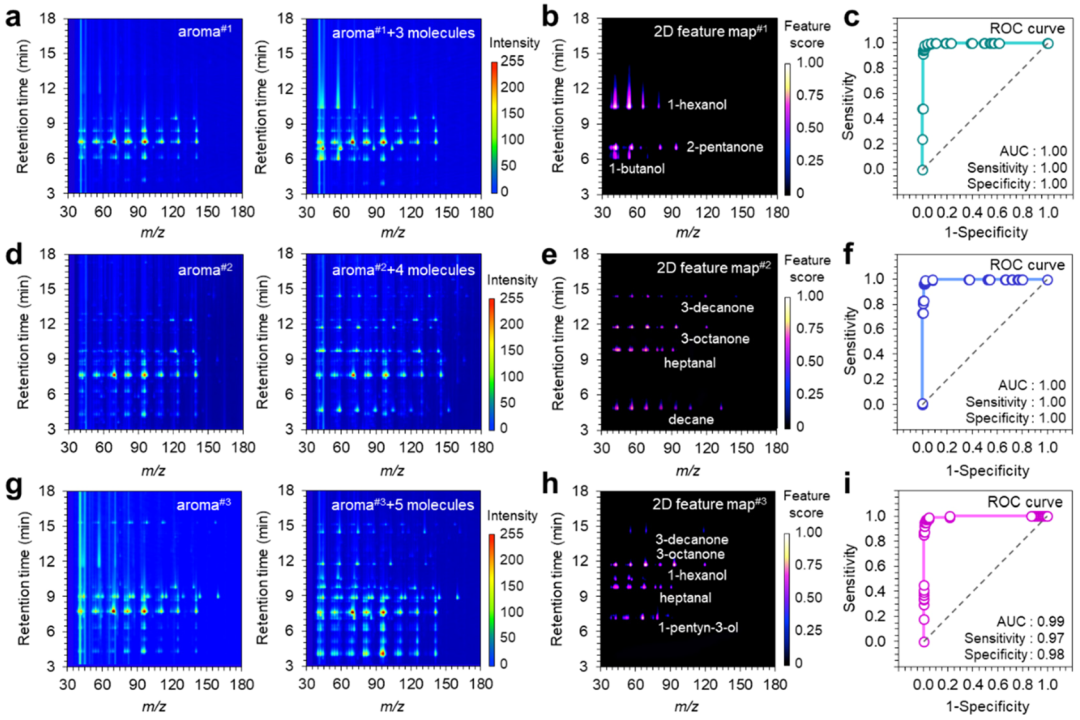


Figure 2. (a,d,g) Regulated 2D MS maps, (b,e,h) 2D feature score maps, and (c,f,i) receiver operating characteristic (ROC) curves of classifiers for (a−c) bergamot organic essential oil——aroma$^{\#1}$, (d−f) lavender essential oil——aroma$^{\#2}$, and (g−i) blend essential oil——aroma$^{\#3}$ in comparison 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.

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-value obtained in t-test of the detected peaks with the number of examined samples.

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

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX

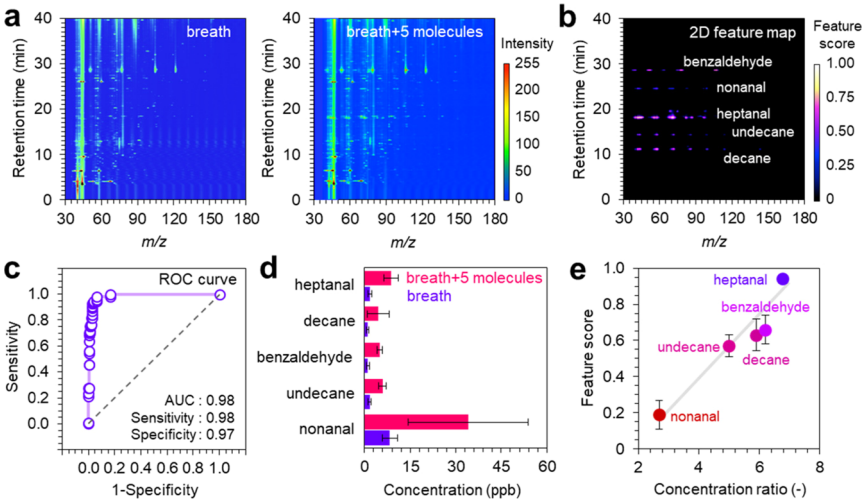


Figure 3. (a) Regulated 2D MS maps, (b) 2D feature score map, and (c) ROC curve of classifiers for breath vs breath + molecule additives, respectively. For the regulated 2D MS maps, the one of original breath is shown in the left and the other with molecule additives is shown in the right. (d) Comparisons 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.

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 well-known 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). 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

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX

S9 and S10). We found that the number of false negative features increased when using the 2D MS map with a linear-scale 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.

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.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c03163.

Workflow details of *NPFimg*; effect of Gaussian filtering on signal intensity; details of position alignment and the standard deviation of the peak positions on 2D MS map; full scale 2D MS map; comparison of the molecular fragment signals in raw MS data and 2D feature score map; optimization of the *XCMS* parameter setting; performances of *NPFimg* and *XCMS* on signal acquisition and feature identification; 2D feature score map based on the random forest algorithm; and applicability of *NPFimg* to untargeted metabolomics (PDF)

AUTHOR INFORMATION

Corresponding Authors

Kazuki Nagashima − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; Japan Science and Technology Agency (JST), PRESTO, Saitama 332-0012, Japan;* orcid.org/0000-0003-0180-816X; Email: kazu-n@g.ecc.u-tokyo.ac.jp

Takeshi Yanagida − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; Interdisciplinary Graduate School of Engineering Sciences and Institute for Materials Chemistry and Engineering, Kyushu University, Fukuoka 816-8580, Japan;* orcid.org/0000-0003-4837-5701; Email: yanagida@g.ecc.u-tokyo.ac.jp

Authors

Chaiyanut Jirayupat − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan*

Takuro Hosomi − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; Japan Science and Technology Agency (JST), PRESTO, Saitama 332-0012, Japan;* orcid.org/0000-0002-5649-6696

Tsunaki Takahashi − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan; Japan Science and Technology Agency (JST), PRESTO, Saitama 332-0012, Japan;* orcid.org/0000-0002-2840-8038

Wataru Tanaka − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

Benjarong Samransuksamer − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

Guozhu Zhang − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX

Jiangyang Liu − *Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

Masaki Kanai − *Institute for Materials Chemistry and Engineering, Kyushu University, Fukuoka 816-8580, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c03163

Author Contributions

C.J. and K.N. designed this work and performed sample preparation/characterization. C.J. constructed the algorithms of *NPFimg* and improved by discussing with K.N., T.H., T.T., W.T., B.S., G.Z., J.L., M.K., and T.Y. C.J. and K.N. wrote this manuscript, and T.H., T.T., W.T., M.K., and T.Y. improved the manuscript. All authors approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by PRESTO Program of Japan Science and Technology Corporation (JST) (No. JPMJPR19J7) and KAKENHI (Nos. JP18H01831, JP18KK0112, JP18H05243, JP20H02208, and JP20F20048). K.N., T.H., T.T., and Y.T. were supported by CREST (No. JPMJCR19I2) and JST Mirai R&D. K.N. was supported by the Research Program for CORE lab of Dynamic Alliance for Open Innovation Bridging Human, Environment and Materials in Network Joint Research Center for Materials and Devices. K.N. acknowledges JACI Prize for Encouraging Young Researcher. This work was partly performed under the Cooperative Research Program of "Network Joint Research Center for Materials and Devices" and the MEXT Project of "Integrated Research Consortium on Chemical Sciences."

■ 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.; Orešič, M. *BMC 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.; Pujos-Guillot, 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.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published on October 27, 2021. Due to production error, some incorrect words were left in the Results and Discussion section. The corrected version was reposted on October 27, 2021.

https://doi.org/10.1021/acs.analchem.1c03163
Anal. Chem. XXXX, XXX, XXX−XXX