analytical chemistry

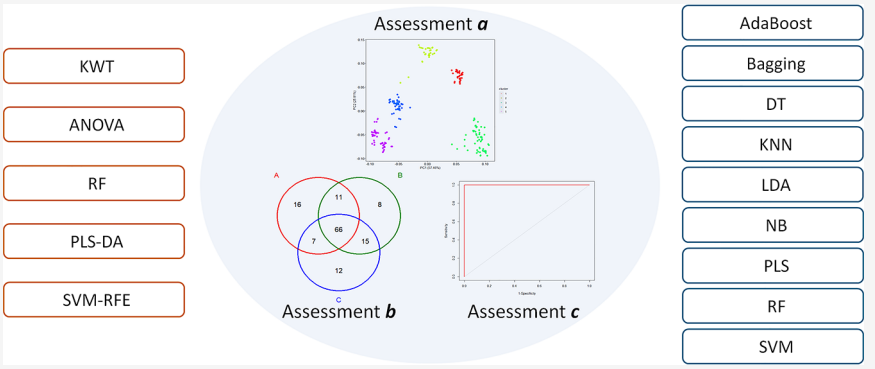
pubs.acs.org/ac Article

Critical Assessment of the Biomarker Discovery and Classification Methods for Multiclass Metabolomics

Qingxia Yang,* Yaguo Gong, and Feng Zhu*

Cite This: *Anal. Chem.* 2023, 95, 5542−5552 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Multiclass metabolomics has been widely applied in clinical practice to understand pathophysiological processes involved in disease progression and diagnostic biomarkers of various disorders. In contrast to the binary problem, the multiclass classification problem is more difficult in terms of obtaining reliable and stable results due to the increase in the complexity of determining exact class decision boundaries. In particular, methods of biomarker discovery and classification have a significant effect on the multiclass model because different methods with significantly varied theories produce conflicting results even for the same dataset. However, a systematic assessment for selecting the most appropriate methods of biomarker discovery and classification for multiclass metabolomics is still lacking. Therefore, a comprehensive assessment is essential to measure the suitability of methods in multiclass classification models from multiple perspectives. In this study, five biomarker discovery methods and nine classification methods were assessed based on four benchmark datasets of multiclass metabolomics. The performance assessment of the biomarker discovery and classification methods was performed using three evaluation criteria: assessment ***a*** (cluster analysis of sample grouping), assessment ***b*** (biomarker consistency in multiple subgroups), and assessment ***c*** (accuracy in the classification model). As a result, 13 combining strategies with superior performance were selected under multiple criteria based on these benchmark datasets. In conclusion, superior strategies that performed consistently well are suggested for the discovery of biomarkers and the construction of a classification model for multiclass metabolomics.

INTRODUCTION

Metabolomics has been recognized as a leading technology that enables simultaneous detection and quantification of subtle variations in metabolites in biological fluids.[1] Metabolome refers to the collection of small-molecule chemical entities. Metabolomics has been applied for the identification of biomarkers for the diagnosis and treatment of disease.[2] Currently, metabolomics has been redefined as a popular technique for identifying biomarkers for discovering active drivers of biological processes.[3] Therefore, metabolomics has been extensively applied in clinical and biomedical studies as a novel and holistic tool for understanding pathophysiological processes involved in disease progression and the identification of diagnostic biomarkers for various disorders.[4]

In metabolomics, the detection of multiclass biosamples is often required for disease diagnoses and clinical applications.[5] There are an increasing number of multiclass ($N > 2$) problems analyzed using metabolomics.[6,7] For example, a multiclass metabolomic study has been applied to reveal the level of bile acids in different cancerous sites,[8] differentiate the presence of succinate in diverse adipose tissues,[9] and discover variations in amino acids of different cell lines.[10] Although there have been various applications for multiclass metabolomics, multiclass studies are intrinsically more difficult for obtaining reliable and stable results than case−control studies.[11]

Received: October 6, 2022
Accepted: March 8, 2023
Published: March 21, 2023

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

Table 1. Key Descriptions of Each Method for Biomarker Discovery and Classification Based on Previous Publications[a]

methods	abbr.	descriptions
(A) biomarker discovery methods in multiclass metabolomics		
Kruskal–Wallis test	KWT	KWT is used to determine the significant differences between the medians of two or more independent groups[38]
one-way ANOVA	ANOVA	ANOVA compares the means of independent groups to determine statistical evidence of population means[45]
partial least squares-discriminant analysis	PLS-DA	PLS-DA is a popular machine learning tool and a supervised feature selector[39]
variable selection from random forests	RF	RF is a machine learning algorithm and it combines the output of multiple decision trees to reach a single result.[40]
support vector machine-recursive feature elimination	SVM-RFE	SVM-RFE can remove relatively insignificant feature variables to achieve higher classification performance[46]
(B) classification methods in multiclass metabolomics		
AdaBoost	AdaBoost	AdaBoost has a high accuracy by focusing on misclassified samples and generating a relatively good model[47]
bagging	bagging	bagging is used to improve accuracy and make the model more generalize by reducing the variance[41]
decision trees	DT	DT is a reliable and effective technique and provides high classification accuracy with the gathered knowledge[48]
K-nearest neighbor	KNN	KNN is one kind of the classical and popular classification approaches, which only need to tune one parameter k[49]
linear discriminant analysis	LDA	LDA is a supervised classifier by creating multiple linear discrimination functions to distinguish different classes[50]
native Bayes	NB	NB takes advantage of probability theory and Bayes' Theorem to the multiclass classification problem[51]
partial least squares	PLS	PLS extracts a set of latent factors and predicts dependent variables using decomposition of independent variables[43]
random forest	RF	RF uses a large series of decision trees with low reciprocal correlation and can model nonlinear relations for dirty data[52]
support vector machine	SVM	SVM is a supervised method that constructs a hyperplane to separate different groups for the given data set[44]

[a]Abbreviation (abbr.) was assigned to indicate each method in the manuscript.

It has been reported that multiclass metabolomic studies are much more complicated due to the complexity of determining exact class decision boundaries.[12,13] In particular, the steps of biomarker discovery and classification have a significant effect on the final results because there are significantly varied theories underlying different methods.[14,15] For metabolomics, various powerful and attractive machine learning techniques exist to analyze complex multiclass data.[11] Fewer than five methods can be applied for biomarker discovery, and fewer than nine classification methods are widely used for constructing classification models. Because of the significantly varied theories underlying each biomarker discovery and classification method, different methods can produce different results even for the same metabolic data.[16,17] Therefore, it is necessary to distinguish the best-performing method from the others for a given dataset. However, systematic assessments of biomarker discovery and classification methods in multiclass metabolomics are lacking. Moreover, a single criterion is insufficient to assess the suitability of those methods. Collective consideration of multiple criteria is recommended for comprehensive assessment from different perspectives. Taken together, a comprehensive assessment using multiple criteria is necessary to evaluate the suitability of biomarker discovery and classification methods.

In this study, a critical assessment of strategies combining biomarker discovery methods and classification methods was performed for multiclass metabolomics. First, different methods (five biomarker discovery methods and nine classification methods) widely used in multiclass metabolomics were collected. Second, four benchmark metabolomic datasets were applied to assess these methods. Third, a comprehensive assessment was performed using three criteria from different perspectives, including assessment α (cluster analysis of sample grouping), assessment β (biomarker consistency in multiple subgroups), and assessment γ (accuracy in the classification model). As a result, 13 combinations of biomarker discovery methods and classification methods were identified as exhibiting superior performance based on the critical assessment. Overall, this study offers metabolic strategies that perform consistently well in constructing a stable classification model for biological problems.

MATERIALS AND METHODS

Collection of Biomarker Discovery Methods and Classification Methods. Five biomarker discovery methods and nine classification methods were included in this study, all widely used in the applications of multiclass metabolomics. In Table 1A, the key descriptions of each biomarker discovery method are summarized based on previous publications, including the Kruskal–Wallis Test (KWT), one-way analysis of variance (ANOVA), partial least squares-discriminant analysis (PLS-DA), variable selection from Random Forest (RF), and support vector machine-recursive feature elimination (SVM-RFE). In Table 1B, the key descriptions of each classification method are summarized based on previous publications, including AdaBoost, bagging, decision trees (DT), K-nearest neighbor (KNN), linear discriminant analysis (LDA), Native Bayes (NB), partial least squares (PLS), Random Forest (RF), and support vector machine (SVM). Detailed information on these 14 methods is presented in the Supplementary Methods. An abbreviation (Table 1) was assigned to each biomarker discovery and classification method correspondingly throughout the manuscript. For example, the combined strategy was depicted as RF + KNN when RF and KNN were applied to discover biomarkers and construct a classification model, respectively. In total, 45 strategies

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

Table 2. Four Benchmark Datasets Used for the Comprehensive Assessment of Biomarker Discovery and Classification Methods in This Study

dataset ID	no. of classes	no. of samples	no. of metabolites in raw data	no. of metabolites in normalized data	type	refs
ST000584 positive mode	5	210	15,447	753	zebrafish	20
ST000584 negative mode	5	210	8781	609	zebrafish	20
ST000880 positive mode	4	47	11,265	6259	mouse	21
ST000880 negative mode	4	47	10,677	4448	mouse	21

combine one biomarker discovery method and one classification method for biomarker discovery and classification, which were comprehensively evaluated for multiclass metabolomics in this study.

Collection of Benchmark Datasets for Multiclass Metabolomics. To assess the performance of all strategies for biomarker discovery and classification in multiclass metabolomics, four benchmark datasets were collected from the public database Metabolomics Workbench (www.metabolomicsworkbench.org).[18] Detailed information on these four datasets is shown in Table 2. The processed dataset was obtained after data preprocessing, including data filtering, imputation for missing values, data transformation, and data normalization from the raw data.[19] The metabolites in all samples were filtered using the 80% rule, and metabolites with >80% missing values were removed from consideration. The remaining missing values of metabolites were imputed using the KNN (K-nearest neighbor imputation) method. After imputation, the dataset was processed using the data transformation (log transformation) method and data normalization (Pareto scaling) method. For dataset ST000584, there were five classes in 210 samples detected in zebrafish.[20] In this dataset, two subsets were named ST000584 positive mode and ST000584 negative mode based on ion mode. For the dataset of ST000584 positive mode, there were 15,447 metabolites in the raw data, and 753 metabolites remained after data preprocessing. For the dataset of ST000584 negative mode, there were 8781 and 609 metabolites in the raw dataset and dataset after data preprocessing, respectively. For dataset ST000880, there were four classes in 47 samples detected in the mouse.[21] Two subsets in this dataset were named ST000880 positive mode and ST000880 negative mode based on ion mode. In the raw datasets, there were 11,265 and 10,677 metabolites for the datasets of ST000880 positive mode and ST000880 negative mode, respectively. After data preprocessing, there were 6259 and 4448 metabolites for the datasets of ST000880 positive mode and ST000880 negative mode, respectively.

Collection of Multiple Criteria for Comprehensive Assessment. *Assessment a: Cluster Analysis of Sample Grouping.* In the multiclass metabolomics analysis, cluster analysis of sample grouping was performed using the *K*-means plot (*K* is the number of classes in the studied dataset). First, differential markers of multiclass metabolomics data were identified using a specific biomarker discovery method. Second, *K*-means clustering was applied to describe the differentiation among different sample groups.[22] An obvious differentiation in this clustering represented the clear separation among different classes for the differential markers. Therefore, the biomarker discovery method used for identifying markers was considered to be performing well. Third, a well-established measure (*purity*) was calculated based on eq 1 and selected to assess the clustering of different classes.[23,24]

$$purity = \sum_{i=1}^{K} \frac{1}{N} \max_{j}(n_i^j) \tag{1}$$

The dataset contains N data objects, K denotes the number of clusters, and n_i^j is the number of objects in the ith cluster belonging to the jth category. If *purity* was higher, the clustering result was more accurate. The worst clustering outcome has a *purity* close to 0, while the perfect clustering result has a *purity* value close to 1.

Assessment b: Biomarker Consistency in Multiple Subgroups. For the same research issue, the low reproducibility of biomarkers identified in different subsets can raise doubt about the consistency of the result.[25] The reason for this low reproducibility of biomarkers is attributed to the inappropriateness of the methods for biomarker discovery.[26,27] Thus, the reproducibility of biomarkers discovered in different subsets is regarded as an essential criterion for assessing the performance of biomarker discovery methods.[28,29] For this criterion, a multiclass dataset was divided into three different sub-datasets by random sampling. Here, stratified random sampling was applied, which involved the division of all samples into different strata (multiple classes), and samples were selected randomly from each stratum. Therefore, the input data can be divided into construct three subgroups using stratified random sampling. Second, the differential metabolic markers were identified from each subgroup using a specific biomarker discovery method. There were three sets of biomarkers in three subgroups using each biomarker discovery method. Third, a powerful measure, relative weighted consistency (*CWrel*), was applied to quantitatively assess the consistency of different sets of biomarkers in three subgroups based on eq 2.[30] It has been reported that *CWrel* is a powerful measure for biomarker consistency in multiple subgroups from an overall perspective to avoid the subset-size-biased problem.[30]

$$CWrel(S, Y) = \frac{|Y|\left(N - D + \sum_{f \in Y}(F_f - 1)\right) - N^2 + D^2}{|Y|(H^2 + n(N - H) - D) - N^2 + D^2} \tag{2}$$

$Y = \{f_1, ..., f_{|Y|}\}$ is the set of all features of size $|Y|$, $S = \{S_1, ..., S_n\}$ is a system of n feature subsets, and F_f denotes the number of occurrences of feature $f \in Y$ in system S. N denotes the total number of occurrences of any feature in system S. This study denotes $D = N \% |Y|$ and $H = N \% n$ for simplicity. The *CWrel* value is between 0 and 1. If the *CWrel* value was close to 1, it indicated the highest consistency of the biomarkers discovered in multiple subgroups.

Assessment c: Accuracy in Classification Model. An important goal of multiclass metabolomics is to identify and validate a set of biomarkers that can be applied to classify multiple classes.[31] First, the differential markers were identified using a biomarker discovery method for the studied multiclass

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

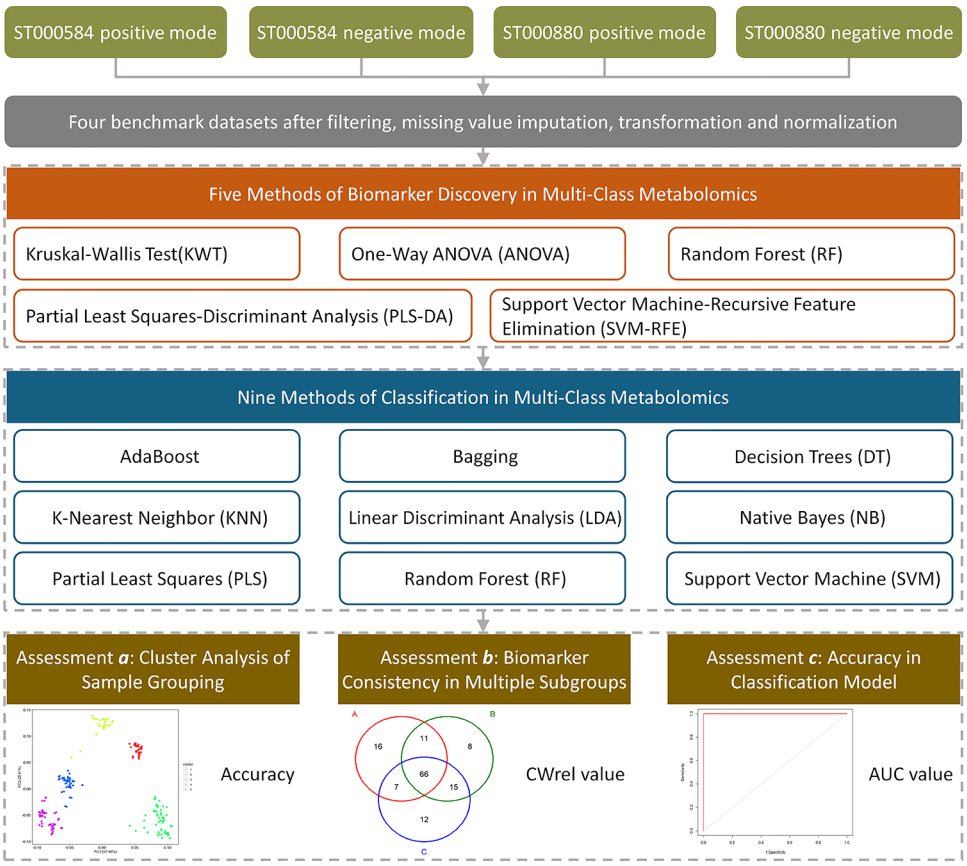


Figure 1. Detailed flowchart in this study included the collection of benchmark datasets for multiclass metabolomics, the collection of biomarker discovery methods and classification methods, and the comprehensive assessment using multiple criteria.

metabolic dataset. Second, using these identified metabolic markers, a classification model was constructed by a classification method. Third, the *AUC* (area under the curve) value of the *ROC* (receiver operating characteristic) curve in the classification model was calculated using the *multiROC* R package.[32] The classification accuracy of the model constructed using the combination of a biomarker discovery method and a classification method was assessed using the *AUC* value.[33,34] The *one-vs-rest* strategy was used in the multiclass classification problem by splitting it into one binary classification problem per class. The average *AUC* value of all binary classification models was applied in the multiclass metabolic dataset. If a classifier achieves high classification performance, the *AUC* value is large (close to 1).

Comprehensive Assessment for Biomarker Discovery and Classification Methods. The detailed flowchart for this study is shown in Figure 1. Five biomarker discovery methods and nine classification methods were assessed using the collective benchmark datasets (Table 2). The comprehensive assessment was performed using representative measures under multiple criteria from multiple perspectives.[16] In particular, the *purity*, *CWrel*, and *AUC* values were measured for Assessment ***a***, Assessment ***b***, and Assessment ***c***, respectively. To assess the robustness of different models, the influence of the size of biomarkers should be examined during the assessment of biomarker discovery and classification methods. Eleven sets with different sizes of biomarkers (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) were generated repetitively. First, *purity* and *CWrel* values were applied to assess the performance of five biomarker discovery methods based on four benchmark datasets. Under each criterion, the *purity* or *CWrel* values for five methods were used to construct five 11-dimensional vectors for each benchmark dataset. Second, the *AUC* value was applied to assess the performance of 45 strategies by combining the biomarker discovery method and classification method. Under this criterion, the *AUC* values for 45 strategies were applied to forty-five 11-dimensional vectors for each benchmark dataset. Third, hierarchical clustering was used to measure the relationship among different methods based on 11-dimensional vectors using R language. In the hierarchical clustering, the Manhattan distance was applied to seek the relationship between any two methods, and Ward's minimum variance method was applied to reduce total within-cluster variance to the maximum extent.[35] The iTOL (Interactive Tree Of Life)[36] tool was applied to draw the graph illustrating the relationship of different methods. As a result, the strategies that performed consistently well were identified and used to construct a stable and reliable classification model for multiclass metabolomics.

RESULTS AND DISCUSSION

Assessment of Biomarker Discovery Methods Using Cluster Analysis of Sample Grouping. To ensure the systematic assessment of biomarker discovery methods, four multiclass metabolic benchmarks were collected and named ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode in Table 2. In these benchmarks, the number of samples varied from dozens to hundreds, and the number of metabolites varied from hundreds to thousands. The various sizes of these

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

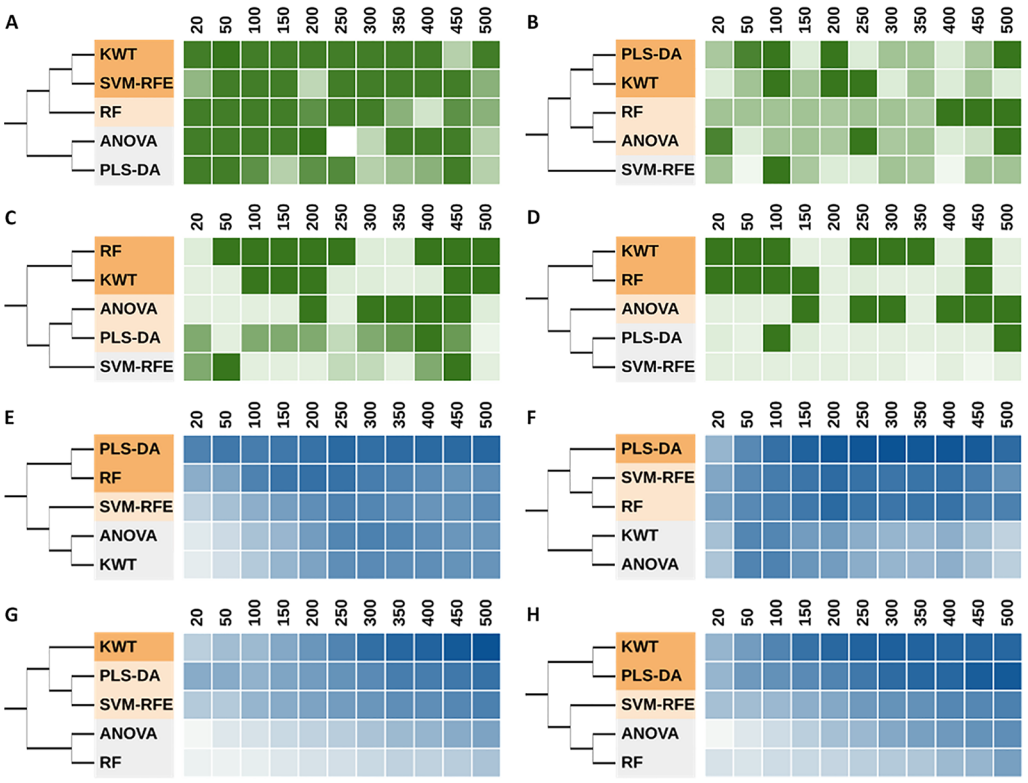


Figure 2. Clusters were performed for biomarker discovery methods assessed by assessment ***a*** (cluster analysis of sample grouping) and assessment ***b*** (biomarker consistency in multiple subgroups). Based on five 11-dimensional vectors for different sizes (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) of biomarkers, *purity*, and *CWrel* values were applied under assessment ***a*** and assessment ***b***, respectively. Clusters of *purity* values were performed using the dataset of (A) ST000584 positive mode, (B) ST000584 negative mode, (C) ST000880 positive mode, and (D) ST000880 negative mode. Clusters of *CWrel* values were performed using the dataset of (E) ST000584 positive mode, (F) ST000584 negative mode, (G) ST000880 positive mode, and (H) ST000880 negative mode. Each cell can represent *purity* or *CWrel* value for one biomarker discovery method. The methods with superior, good, and poor performance are colored orange, light orange, and gray, respectively.

benchmarks can support the systematic evaluation of these biomarker discovery methods. In particular, all studied methods were evaluated using cluster analysis of sample grouping under assessment ***a***, which was quantitatively measured using the *purity* value. As shown in Supplementary Table S1, the performance of all biomarker discovery methods based on four benchmarks was assessed by cluster analysis of sample grouping under assessment ***a***. For each benchmark dataset, the *purity* values among different sizes (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) of biomarkers were used to construct five 11-dimensional vectors for five biomarker discovery methods. These five 11-dimensional vectors using each benchmark were applied to assess the relationships among the performance of different methods.

These relationships were studied based on hierarchical clustering using the 11-dimensional vectors. The results of hierarchical clustering among the five biomarker discovery methods based on the ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode are shown in Figure 2A–D, respectively. The methods with similar performance were clustered together to discover consistently well-performing methods irrespective of the number of biomarkers. Methods with superior, good, and poor performance were colored orange, light orange, and gray, respectively. The results indicate that the methods exhibiting superior performance conflict when using different benchmark datasets. In Figure 2A, the KWT, SVM-RFE, and RF were identified as well-performing methods using the dataset of ST000584 positive mode. In Figure 2B, PLS-DA, KWT, RF, and ANOVA were identified as well-performing methods using the dataset of ST000584 negative mode. In Figure 2C, RF, KWT, ANOVA, and PLS-DA were identified as well-performing methods using the dataset of ST000880 positive mode. As shown in Figure 2D, the KWT, RF, and ANOVA were identified as well-performing methods using the dataset of ST000880 negative mode. Among these well-performing methods using these benchmarks, KWT and RF were consistently well-performing in four benchmark datasets under the criterion of cluster analysis of sample grouping. Similarly, ANOVA, PLS-DA, and SVM-RFE were also identified from three datasets, two datasets, and one dataset, respectively. Using raincloud plots, the *purity* values for these methods are illustrated in Figure 3A–D for the datasets of ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. The *purity* values of KWT in the raincloud plot were consistently higher than those of the other methods from the four raincloud plots.

Assessment of Biomarker Discovery Methods by Biomarker Consistency in Multiple Subgroups. Based on four benchmark datasets of multiclass metabolomics (Table 2), the performance of all biomarker discovery methods was evaluated using biomarker consistency in multiple subgroups.[37] Under this criterion, the performance of all methods was quantitatively measured using the *CWrel* value. As illustrated in Supplementary Table S2, the *CWrel* values of biomarker

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

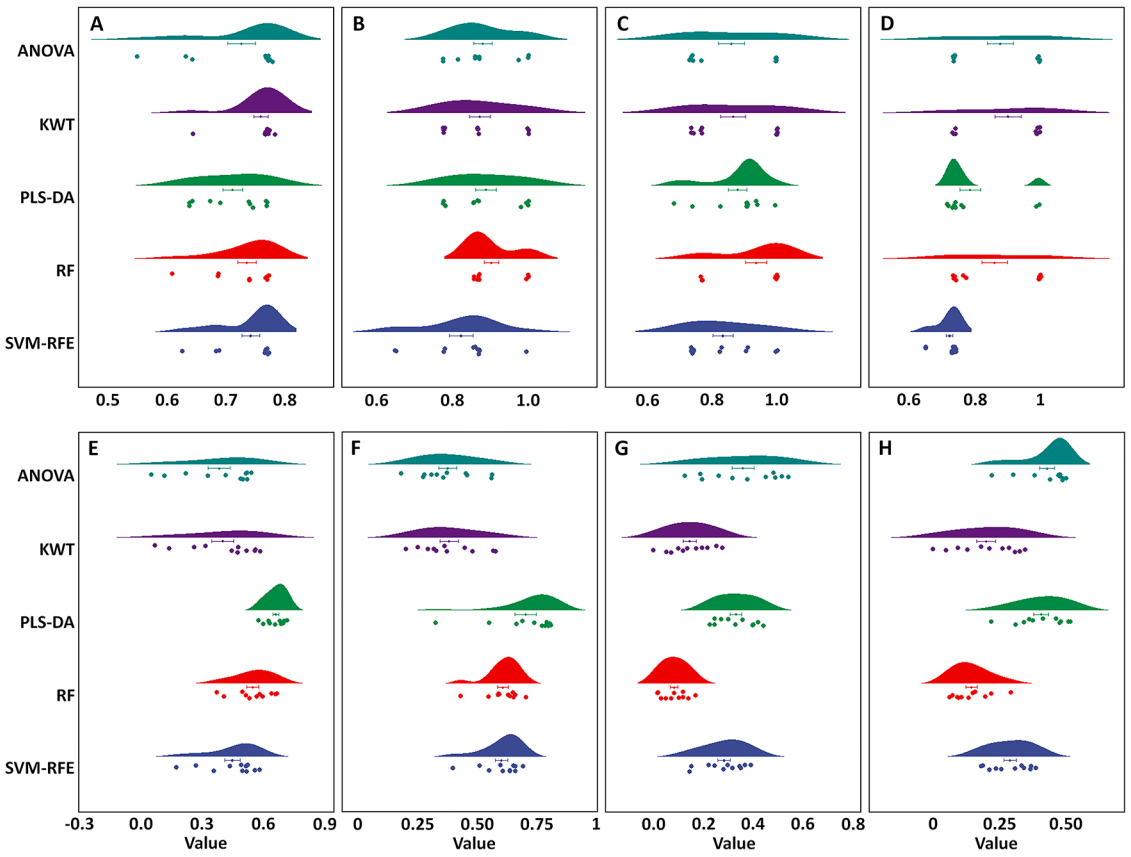


Figure 3. Raincloud plots for the assessment performance of biomarker discovery methods. The performance was assessed using *purity* values under assessment ***a*** (cluster analysis of sample grouping) using the dataset of (A) ST000584 positive mode, (B) ST000584 negative mode, (C) ST000880 positive mode, and (D) ST000880 negative mode. The performance was assessed using *CWrel* values under assessment ***b*** (biomarker consistency in multiple subgroups) using the datasets of (E) ST000584 positive mode, (F) ST000584 negative mode, (G) ST000880 positive mode, and (H) ST000880 negative mode.

discovery methods based on four benchmarks are fully provided. For each benchmark dataset, the *CWrel* values among different sizes (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) of biomarkers were applied to construct five 11-dimensional vectors for five biomarker discovery methods. These five 11-dimensional vectors using each benchmark were applied to assess the relationships among the performance of different methods under assessment ***b***.

Using hierarchical clustering, the relationships among the five biomarker discovery methods are shown in Figure 2E−H for the datasets of ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. The methods with similar performance were clustered together, helping to discover the methods irrespective of the different number of biomarkers. Methods with superior, good, and poor performance were colored orange, light orange, and gray, respectively. As a result, the well-performing methods assessed by biomarker consistency in multiple subgroups conflict when using different benchmark datasets. In Figure 2E, PLS-DA, RF, and SVM-RFE were identified as well-performing methods using the dataset of ST000584 positive mode. In Figure 2F, PLS-DA, SVM-RFE, and RF were identified as well-performing methods using the dataset of ST000584 negative mode. In Figure 2G, KWT, PLS-DA, and SVM-RFE were identified as well-performing methods using the dataset of ST000880 positive mode. In Figure 2H, KWT, PLS-DA, and SVM-RFE were identified as well-performing methods using the dataset of ST000880 negative mode. Among these well-performing methods under the criterion of biomarker consistency in multiple subgroups, PLS-DA and SVM-RFE were consistently well-performing in four benchmark datasets. Similarly, KWT and RF were identified simultaneously in two benchmark datasets. Using raincloud plots, the *CWrel* values for each method are illustrated in Figure 3E−H for ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. The *CWrel* values of PLS-DA in the raincloud plot were consistently higher than those of other methods from the four raincloud plots.

Assessment of Biomarker Discovery and Classification Methods by Classification Accuracy. The performance of strategies formed by combining the biomarker discovery method and classification method was assessed using the accuracy of the classification model using four benchmark datasets (Table 2). Under assessment ***c***, *AUC* values were applied for quantitative measurement of classification accuracy. As shown in Supplementary Table S3, the performance of all combined strategies is fully provided based on the accuracy in the classification model using four benchmarks. For each benchmark dataset, the *AUC* values among different sizes (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) of biomarkers were used to construct forty-five 11-dimensional vectors for 45 strategies combining the biomarker discovery method and classification method. These forty-five 11-dimensional vectors using each benchmark were applied to assess the relationships among the performance of

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

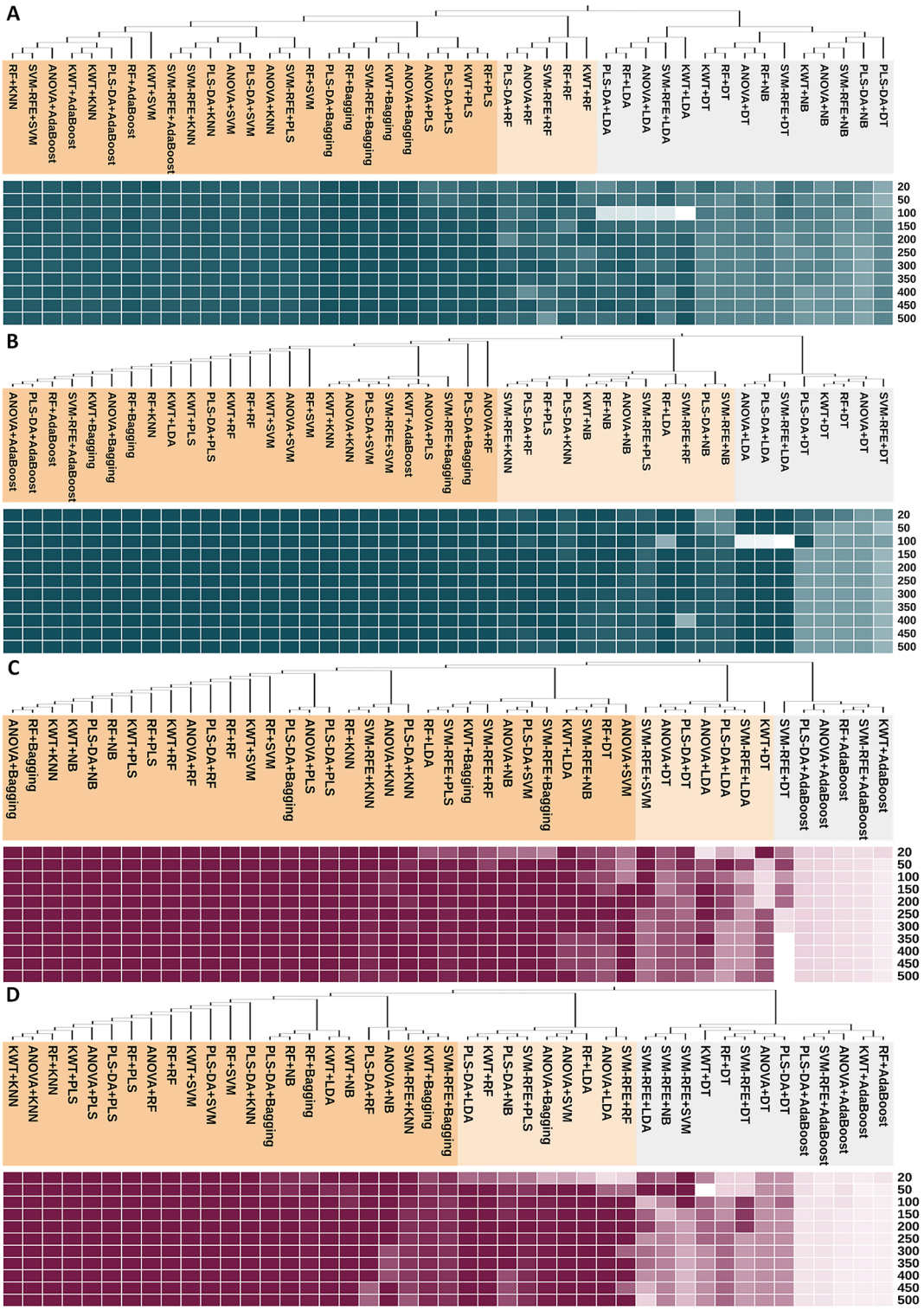


Figure 4. Clusters of performance assessment were performed for 45 strategies by combining the biomarker discovery method and classification method under assessment *c* (accuracy in classification model). *AUC* values were applied under assessment *c* based on five eleven-dimensional vectors for different sizes (top 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) of biomarkers. Clusters of *AUC* values were performed based on the dataset of (A) ST000584 positive mode, (B) ST000584 negative mode, (C) ST000880 positive mode, and (D) ST000880 negative mode. Each cell can represent the *AUC* value for one combining strategy. The strategies with superior, good, and poor performance are colored orange, light orange, and gray, respectively.

different strategies. Using hierarchical clustering, the relationships among 45 combining strategies are shown in Figure 4A–D for the datasets of ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. For plots of hierarchical clustering, strategies with superior, good, and poor performance were colored orange, light orange, and gray, respectively. The intersection of the superior strategies using four different

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

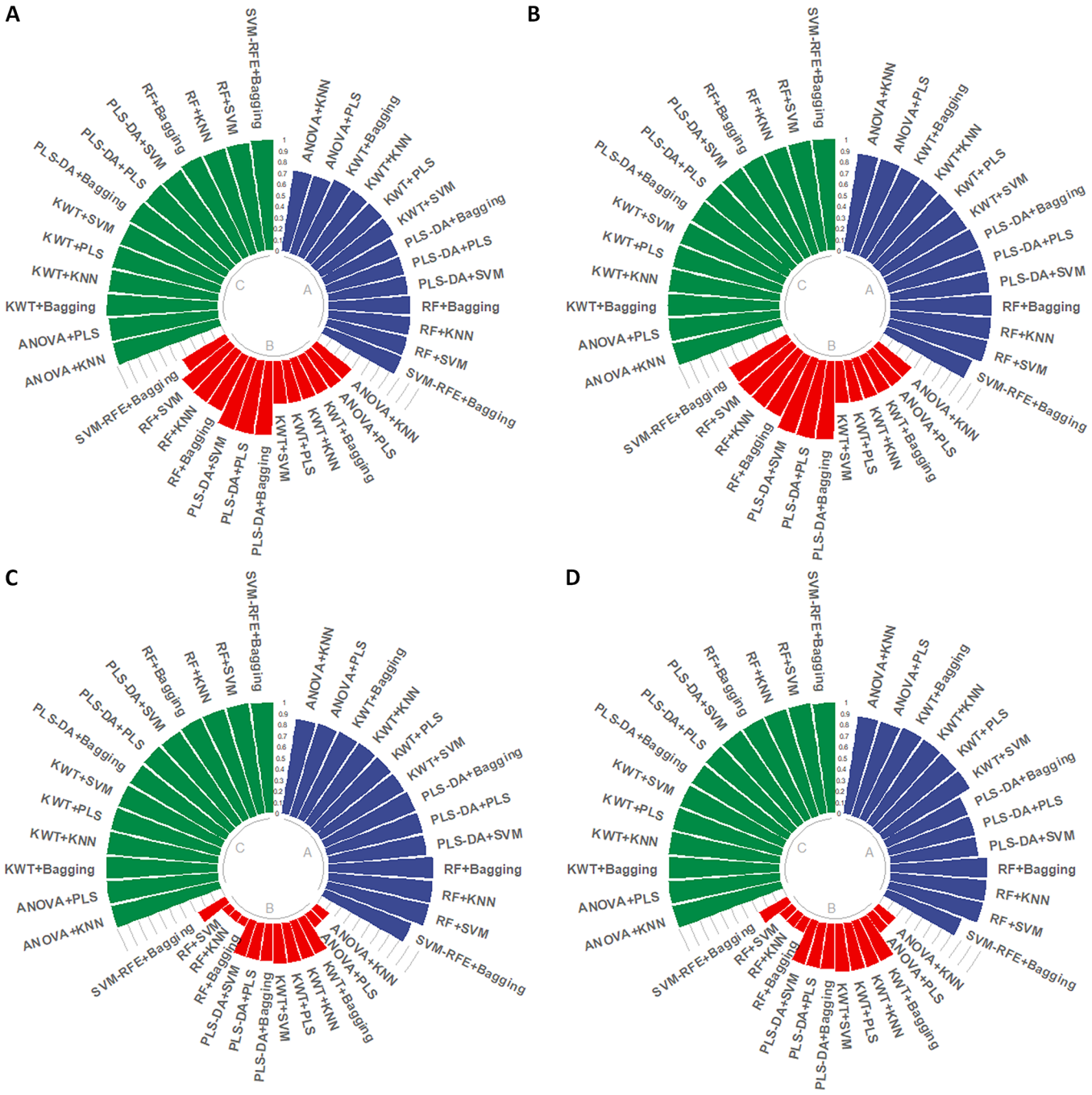


Figure 5. Circular bar plots of 13 strategies with superior performance by combining biomarker discovery methods and classification methods assessed by assessment ***c*** (accuracy in classification model). Each strategy of *purity*, *CWrel*, and *AUC* value is colored blue, red, and green, respectively. Circular bar plots were generated using the dataset of (A) ST000584 positive mode, (B) ST000584 negative mode, (C) ST000880 positive mode, and (D) ST000880 negative mode.

benchmarks was shown to be consistently well-performing strategies.

Discovering Consistently Well-Performing Classification Models under All Criteria. Using four benchmarks, there were 13 strategies with superior performance, combining the biomarker discovery method and classification method. These strategies included ANOVA + KNN, ANOVA + PLS, KWT + bagging, KWT + KNN, KWT + PLS, KWT + SVM, PLS-DA + bagging, PLS-DA + PLS, PLS-DA + SVM, RF + bagging, RF + KNN, RF + SVM, and SVM-RFE + bagging. Except for ANOVA, other methods of biomarker discovery including KWT, PLS-DA, RF, and SVM-RFE were identified as well-performing methods under assessment ***a*** and assessment ***b***. Using circular bar plots, detailed information on these superior strategies using the measured metric under each assessment criterion is illustrated in Figure 5. These combining strategies under three assessment criteria are shown in Figure 5A−D for the dataset of ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. These strategies using *purity*, *CWrel*, and *AUC* values under assessment ***a***, assessment ***b***, and assessment ***c*** were colored blue, red, and green, respectively.

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

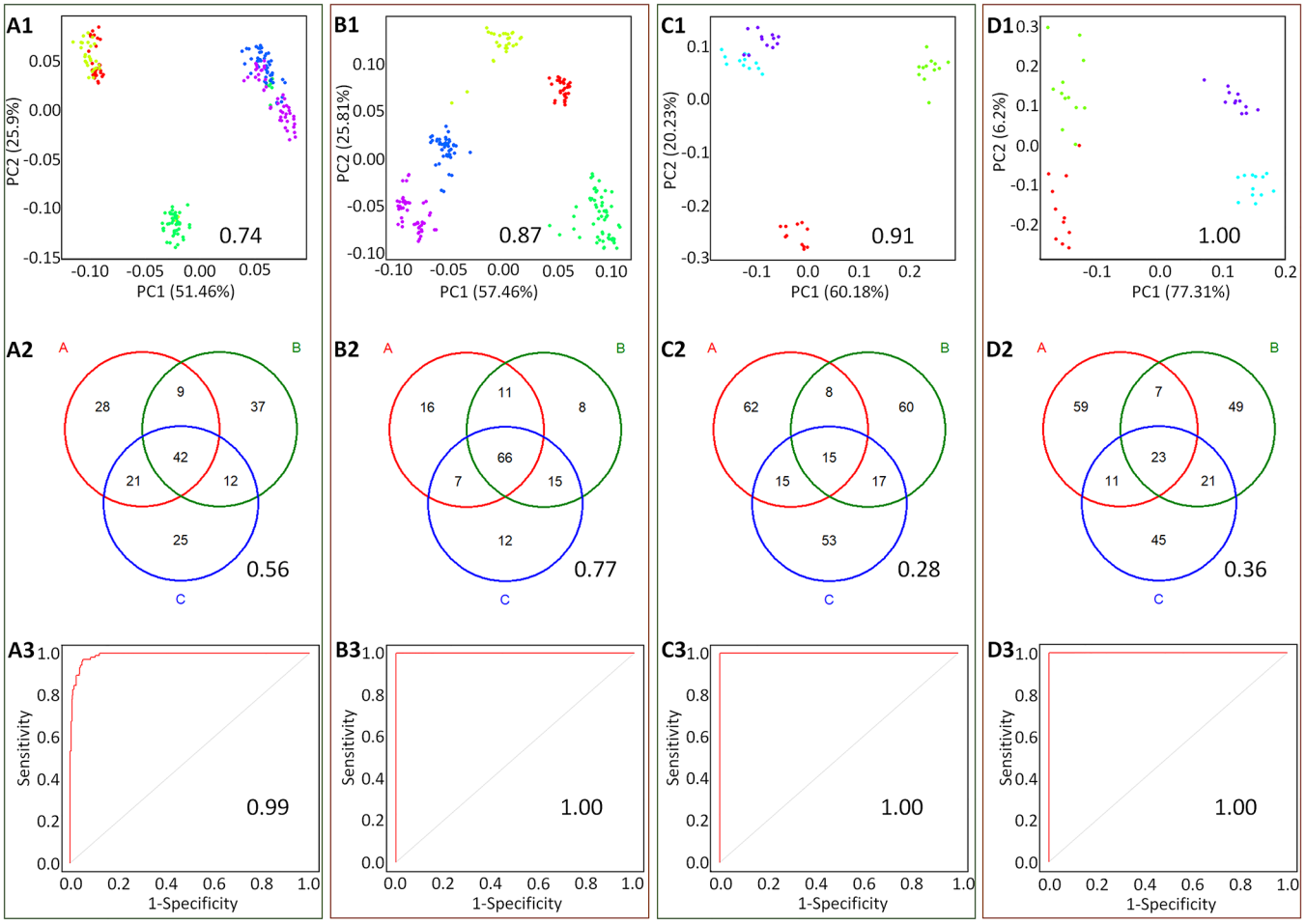


Figure 6. Results of assessment performance for the combining strategy (PLS-DA + bagging) under three criteria using benchmark datasets. For the dataset of (A) ST000584 positive mode, the plots of results assessed using assessment ***a*** (cluster analysis of sample grouping), assessment ***b*** (biomarker consistency in multiple subgroups), and assessment ***c*** (accuracy in classification model) are shown in A1, A2, and A3, respectively. For the dataset of (B) ST000584 negative mode, the plots of results assessed using assessment ***a***, assessment ***b***, and assessment ***c*** are shown in B1, B2, and B3, respectively. For the dataset of (C) ST000880 positive mode, the plots of results assessed using assessment ***a***, assessment ***b***, and assessment ***c*** are shown in C1, C2, and C3, respectively. For the dataset of (D) ST000880 negative mode, the plots of results assessed using assessment ***a***, assessment ***b***, and assessment ***c*** are shown in D1, D2, and D3, respectively.

For these benchmark datasets, the sample number varied from dozens to 100, including four or five sample classes, and the number of metabolites varied from 600 to 6000. The diversity of the benchmarks ensured the comprehensive assessment of the combined strategies.

From the blue areas of Figure 5, all 13 combining strategies performed consistently well under assessment ***a*** (cluster analysis of sample grouping). The range of *purity* values varied from 0.71 to 0.76 for the dataset of ST000584 positive mode, and the *purity* values were greater than 0.80 for the dataset of ST000584 negative mode and ST000880 positive mode. For the ST000880 negative mode, the *purity* values of four combining strategies varied from 0.73 to 0.79, including PLS-DA + bagging, PLS-DA + PLS, PLS-DA + SVM, and SVM-RFE + bagging, and the values of others were greater than 0.80. From the red areas of Figure 5, the range of *CWrel* values for all 13 combining strategies was widely distributed from 0.08 to 0.70 under assessment ***b*** (biomarker consistency in multiple subgroups). In particular, three combining strategies, including PLS-DA + SVM, PLS-DA + PLS, and PLS-DA + bagging, performed consistently well with *CWrel* values greater than 0.30. As indicated by the green areas of Figure 5, all 13 combining strategies performed consistently well under assessment ***c*** (accuracy of classification model) with *CWrel* values greater than 0.98 for all benchmarks.

The combining strategy, PLS-DA + bagging, was selected as an example to visualize the performance under three assessment criteria based on four benchmark datasets. The top 100 biomarkers of the PLS-DA model were applied for a comprehensive assessment. The results of the assessment under three criteria are shown in Figure 6A–D for the dataset of ST000584 positive mode, ST000584 negative mode, ST000880 positive mode, and ST000880 negative mode, respectively. In Figure 6A1–D1, samples of different groups were separated in the cluster analysis using the *K*-means algorithm. The *purity* values in cluster analysis of sample grouping were 0.74, 0.87, 0.91, and 1.00 for four benchmarks. In Figure 6A2–D2, the overlap of markers among the three subgroups was 42, 66, 15, and 23 for the four benchmarks. The *CWrel* values under the criterion of biomarker consistency in multiple subgroups were 0.56, 0.77, 0.28, and 0.36 for different benchmark datasets. In Figure 6A3–D3, the *AUC* value in the *ROC* curve under the criterion of accuracy in the classification model was 0.99, 1.00, 1.00, and 1.00 for four different

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552

benchmarks. Therefore, these 13 combined strategies of the biomarker discovery method and classification method consistently performed well under multiple criteria for multiclass metabolomics.

The strategies with superior performance were identified using the comprehensive assessment of three measures based on four different benchmark datasets. Among these superior strategies, three biomarker discovery methods (KWT, PLS-DA, and RF) were discovered in more than three combined strategies. Four classification methods (bagging, KNN, PLS, and SVM) were discovered in more than three combined strategies. The advantages and superior performance of these methods have been reported in many kinds of studies. The performance of KWT was more robust against departures from the assumption of the equality of variance.[38] When the sparsity of the data grows increasingly faster, it is easier for PLS-DA to detect the strong correlation between the signal features for class members.[39] Using the sum of the decisions, RF has consistently lower generalization errors than others.[40] For bagging, the advantage is that multiple weak learners can provide stability, increase accuracy, and avoid overfitting comparing to a single strong learner.[41] KNN has many advantages for the classification task without any assumption for the distribution of a large number of training data.[42] PLS is particularly useful when predictor variables are highly correlated by extracting a set of latent factors.[43] By obtaining a hyperplane with the greatest distance to the nearest training data of any class, SVM is suitable for classification with higher speed and better performance.[44]

This study performed a critical assessment of biomarker discovery and classification methods for multiclass metabolomics. However, there were some limitations. First, more benchmark datasets should be applied to assess the performance of these biomarker discovery and classification methods. Second, three assessment measures (cluster analysis of sample grouping, biomarker consistency in multiple subgroups, and accuracy in classification model) were applied for a comprehensive assessment of the biomarker discovery and classification methods. The level of correspondence between detected biomarkers and spike-in metabolites can be used to assess different methods. However, it is difficult to annotate the biomarkers for the benchmark datasets from the public database in this study. In the future, experimental data including spike-in metabolites are necessary to assess the level of correspondence between detected biomarkers and spike-in metabolites. Third, the superior strategies of combining biomarker discovery and classification methods might be slightly inconsistent due to the unique attributes of the input data. In the future, a computational tool is still needed to identify the appropriate biomarker discovery and classification method for multiclass metabolomics.

CONCLUSIONS

In this study, a critical assessment was performed for biomarker discovery and classification methods in multiclass metabolomics. Five biomarker discovery methods and nine classification methods were assessed based on four benchmark datasets. The comprehensive assessment was performed using three assessment criteria: assessment ***a*** (cluster analysis of sample grouping), assessment ***b*** (biomarker consistency in multiple subgroups), and assessment ***c*** (accuracy in classification model). As a result, 13 strategies combining the biomarker discovery method and classification method were discovered to be strategies with superior performance. In conclusion, this study can provide clues for constructing a classification model for multiclass metabolomics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c04402.

Introduction of biomarker discovery and classification methods in multiclass metabolomics and detailed information on the performance of all combined strategies under three assessment criteria (Tables S1–S3) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Qingxia Yang − *Department of Bioinformatics, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-9607-7026; Email: yangqx@njupt.edu.cn

Feng Zhu − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-8069-0053; Email: zhufeng@zju.edu.cn

Author

Yaguo Gong − *School of Pharmacy, Macau University of Science and Technology, Macau 999078, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c04402

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (62201289), the Natural Science Foundation of Jiangsu Province (BK20210597), and NUPTSF (NY220169).

REFERENCES

(1) Bauermeister, A.; Mannochio-Russo, H.; Costa-Lotufo, L. V.; Jarmusch, A. K.; Dorrestein, P. C. *Nat. Rev. Microbiol.* **2022**, *20*, 143–160.

(2) Xiao, Y.; Ma, D.; Yang, Y. S.; Yang, F.; Ding, J. H.; Gong, Y.; Jiang, L.; Ge, L. P.; Wu, S. Y.; Yu, Q.; Zhang, Q.; Bertucci, F.; Sun, Q.; Hu, X.; Li, D. Q.; Shao, Z. M.; Jiang, Y. Z. *Cell Res.* **2022**, *32*, 477–490.

(3) Rinschen, M. M.; Ivanisevic, J.; Giera, M.; Siuzdak, G. *Nat. Rev. Mol. Cell. Biol.* **2019**, *20*, 353–367.

(4) Talmor-Barkan, Y.; Bar, N.; Shaul, A. A.; Shahaf, N.; Godneva, A.; Bussi, Y.; Lotan-Pompan, M.; Weinberger, A.; Shechter, A.; Chezar-Azerrad, C.; Arow, Z.; Hammer, Y.; Chechi, K.; Forslund, S. K.; Fromentin, S.; Dumas, M. E.; Ehrlich, S. D.; Pedersen, O.; Kornowski, R.; Segal, E. *Nat. Med.* **2022**, *28*, 295–302.

(5) Xia, W.; Zheng, L.; Fang, J.; Li, F.; Zhou, Y.; Zeng, Z.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. *Comput. Biol. Med.* **2022**, *145*, No. 105465.

(6) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. *Nucleic Acids Res.* **2020**, *48*, W436–W448.

(7) Cladiere, M.; Delaporte, G.; Le Roux, E.; Camel, V. *Food Chem.* **2018**, *242*, 113–121.

(8) Lee, C. K.; Jeong, S. H.; Jang, C.; Bae, H.; Kim, Y. H.; Park, I.; Kim, S. K.; Koh, G. Y. *Science* **2019**, *363*, 644–649.

(9) Mills, E. L.; Pierce, K. A.; Jedrychowski, M. P.; Garrity, R.; Winther, S.; Vidoni, S.; Yoneshiro, T.; Spinelli, J. B.; Lu, G. Z.; Kazak, L.; Banks, A. S.; Haigis, M. C.; Kajimura, S.; Murphy, M. P.; Gygi, S. P.; Clish, C. B.; Chouchani, E. T. *Nature* **2018**, *560*, 102−106.
(10) Yan, W.; Wu, X.; Zhou, W.; Fong, M. Y.; Cao, M.; Liu, J.; Liu, X.; Chen, C. H.; Fadare, O.; Pizzo, D. P.; Wu, J.; Liu, L.; Liu, X.; Chin, A. R.; Ren, X.; Chen, Y.; Locasale, J. W.; Wang, S. E. *Nat. Cell Biol.* **2018**, *20*, 597−609.
(11) Yang, Q.; Xu, L.; Tang, L. J.; Yang, J. T.; Wu, B. Q.; Chen, N.; Jiang, J. H.; Yu, R. Q. *Talanta* **2018**, *186*, 489−496.
(12) Rocha, A.; Goldenstein, S. K. *IEEE Trans. Neural Networks Learn. Syst.* **2014**, *25*, 289−302.
(13) Yang, Q.; Li, Y.; Li, B.; Gong, Y. *Comput. Biol. Med.* **2022**, *148*, No. 105956.
(14) Helf, M. J.; Fox, B. W.; Artyukhin, A. B.; Zhang, Y. K.; Schroeder, F. C. *Nat. Commun.* **2022**, *13*, 782.
(15) Yang, Q.; Xing, Q.; Yang, Q.; Gong, Y. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 5054−5064.
(16) Yang, Q.; Hong, J.; Li, Y.; Xue, W.; Li, S.; Yang, H.; Zhu, F. *Briefings Bioinf.* **2020**, *21*, 2142−2152.
(17) Yang, Q.; Wang, Y.; Zhang, S.; Tang, J.; Li, F.; Yin, J.; Li, Y.; Fu, J.; Li, B.; Luo, Y.; Xue, W.; Zhu, F. *Int. J. Mol. Sci.* **2019**, 20.
(18) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44*, D463−D470.
(19) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. *Nat. Protoc.* **2022**, *17*, 129−151.
(20) Dvornikov, A. V.; Wang, M.; Yang, J.; Zhu, P.; Le, T.; Lin, X.; Cao, H.; Xu, X. *J. Mol. Cell. Cardiol.* **2019**, *133*, 199−208.
(21) Fujisaka, S.; Avila-Pacheco, J.; Soto, M.; Kostic, A.; Dreyfuss, J. M.; Pan, H.; Ussar, S.; Altindis, E.; Li, N.; Bry, L.; Clish, C. B.; Kahn, C. R. *Cell Rep.* **2018**, *22*, 3072−3086.
(22) Jacob, S.; Nodzenski, M.; Reisetter, A. C.; Bain, J. R.; Muehlbauer, M. J.; Stevens, R. D.; Ilkayeva, O. R.; Lowe, L. P.; Metzger, B. E.; Newgard, C. B.; Scholtens, D. M.; Lowe, W. L., Jr.; HAPO Study Cooperative Research Group. *Diabetes Care* **2017**, *40*, 911−919.
(23) Jiang, H.; Sohn, L. L.; Huang, H.; Chen, L. *Bioinformatics* **2018**, *34*, 3684−3694.
(24) Huang, S.; Cheng, Y.; Lang, D.; Chi, R.; Liu, G. *PLoS One* **2014**, *9*, No. e90109.
(25) Wang, Y.; Klijn, J. G.; Zhang, Y.; Sieuwerts, A. M.; Look, M. P.; Yang, F.; Talantov, D.; Timmermans, M.; Meijer-van Gelder, M. E.; Yu, J.; Jatkoe, T.; Berns, E. M.; Atkins, D.; Foekens, J. A. *Lancet* **2005**, *365*, 671−679.
(26) Yeung, K. Y.; Bumgarner, R. E. *Genome Biol.* **2003**, *4*, R83.
(27) Yang, Q.; Li, B.; Tang, J.; Cui, X.; Wang, Y.; Li, X.; Hu, J.; Chen, Y.; Xue, W.; Lou, Y.; Qiu, Y.; Zhu, F. *Briefings Bioinf.* **2020**, *21*, 1058−1068.
(28) Student, S.; Fujarewicz, K. *Biol. Direct* **2012**, *7*, 33.
(29) Yang, Q. X.; Wang, Y. X.; Li, F. C.; Zhang, S.; Luo, Y. C.; Li, Y.; Tang, J.; Li, B.; Chen, Y. Z.; Xue, W. W.; Zhu, F. *CNS Neurosci. Ther.* **2019**, *25*, 1054−1063.
(30) Somol, P.; Novovicova, J. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1921−1939.
(31) Peeters, L.; Beirnaert, C.; Van der Auwera, A.; Bijttebier, S.; De Bruyne, T.; Laukens, K.; Pieters, L.; Hermans, N.; Foubert, K. *J. Chromatogr., A* **2019**, *1595*, 240−247.
(32) Shultz, E. K. *Clin. Chem.* **1995**, *41*, 1248−1255.
(33) Grissa, D.; Petera, M.; Brandolini, M.; Napoli, A.; Comte, B.; Pujos-Guillot, E. *Front. Mol. Biosci.* **2016**, *3*, 30.
(34) Yang, Q.; Li, B.; Chen, S.; Tang, J.; Li, Y.; Li, Y.; Zhang, S.; Shi, C.; Zhang, Y.; Mou, M.; Xue, W.; Zhu, F. *J. Proteomics* **2021**, *232*, No. 104023.
(35) Kim, J.; Mouw, K. W.; Polak, P.; Braunstein, L. Z.; Kamburov, A.; Kwiatkowski, D. J.; Rosenberg, J. E.; Van Allen, E. M.; D'Andrea, A.; Getz, G. *Nat. Genet.* **2016**, *48*, 600−606.
(36) Letunic, I.; Bork, P. *Nucleic Acids Res.* **2021**, *49*, W293−W296.
(37) Li, F.; Zhou, Y.; Zhang, Y.; Yin, J.; Qiu, Y.; Gao, J.; Zhu, F. *Briefings Bioinf.* **2022**, 23.
(38) Fan, C.; Zhang, D.; Zhang, C. H. *Biometrics* **2011**, *67*, 213−224.
(39) Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G. *BMC Bioinf.* **2020**, *21*, 2.
(40) Heo, J.; Yoon, J. G.; Park, H.; Kim, Y. D.; Nam, H. S.; Heo, J. H. *Stroke* **2019**, *50*, 1263−1265.
(41) Dudoit, S.; Fridlyand, J. *Bioinformatics* **2003**, *19*, 1090−1099.
(42) Yu, J.; Tao, D.; Wang, M.; Rui, Y. *IEEE Trans. Cybern.* **2015**, *45*, 767−779.
(43) Lee, L. C.; Liong, C. Y.; Jemain, A. A. *Analyst* **2018**, *143*, 3526−3539.
(44) Webb-Robertson, B. J.; Cannon, W. R.; Oehmen, C. S.; Shah, A. R.; Gurumoorthi, V.; Lipton, M. S.; Waters, K. M. *Bioinformatics* **2008**, *24*, 1503−1509.
(45) Jan, S. L.; Shieh, G. *Br. J. Math. Stat. Psychol.* **2014**, *67*, 72−93.
(46) Huang, M. L.; Hung, Y. H.; Lee, W. M.; Li, R. K.; Jiang, B. R. *Sci. World J.* **2014**, *2014*, 795624.
(47) Fu, W.; Yu, S.; Wang, X. *Entropy* **2021**, 23.
(48) Luna, J. M.; Gennatas, E. D.; Ungar, L. H.; Eaton, E.; Diffenderfer, E. S.; Jensen, S. T.; Simone, C. B., 2nd; Friedman, J. H.; Solberg, T. D.; Valdes, G. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 19887−19893.
(49) Yu, Z.; Chen, H.; Liuxs, J.; You, J.; Leung, H.; Han, G. *IEEE Trans. Cybern.* **2016**, *46*, 1263−1275.
(50) Dornaika, F.; Khoder, A. *Neural Networks* **2020**, *127*, 141−159.
(51) Trainor, P. J.; DeFilippis, A. P.; Rai, S. N. *Metabolites* **2017**, *7*, 30.
(52) Yang, L.; Wu, H.; Jin, X.; Zheng, P.; Hu, S.; Xu, X.; Yu, W.; Yan, J. *Sci. Rep.* **2020**, *10*, 5245.

https://doi.org/10.1021/acs.analchem.2c04402
Anal. Chem. 2023, 95, 5542−5552