Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3282−3294

pubs.acs.org/jpr

Evaluation of Multivariate Classification Models for Analyzing NMR Metabolomics Data

Thao Vu,[†] Parker Siemek,[‡] Fatema Bhinderwala,[‡,§] Yuhang Xu,[†,⊥] and Robert Powers*[,‡,§]

[†]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68583-0963, United States

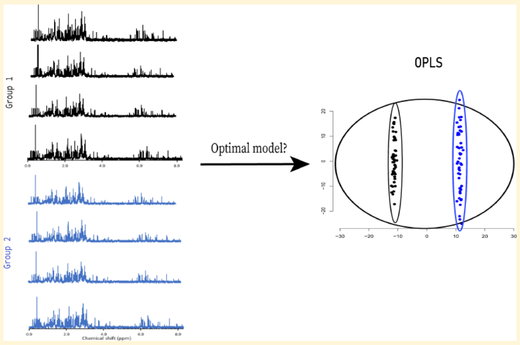
[‡]Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States

[§]Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States

[⊥]Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, Ohio 43403-0001, United States

Supporting Information

ABSTRACT: Analytical techniques such as NMR and mass spectrometry can generate large metabolomics data sets containing thousands of spectral features derived from numerous biological observations. Multivariate data analysis is routinely used to uncover the underlying biological information contained within these large metabolomics data sets. This is typically accomplished by classifying the observations into groups (e.g., control versus treated) and by identifying associated discriminating features. There are a variety of classification models to select from, which include some well-established techniques (e.g., principal component analysis [PCA], orthogonal projection to latent structure [OPLS], or partial least-squares projection to latent structures [PLS]) and newly emerging machine learning algorithms (e.g., support vector machines or random forests). However, it is unclear which classification model, if any, is an optimal choice for the analysis of metabolomics data. Herein, we present a comprehensive evaluation of five common classification models routinely employed in the metabolomics field and that are also currently available in our MVAPACK metabolomics software package. Simulated and experimental NMR data sets with various levels of group separation were used to evaluate each model. Model performance was assessed by classification accuracy rate, by the area under a receiver operating characteristic (AUROC) curve, and by the identification of true discriminating features. Our findings suggest that the five classification models perform equally well with robust data sets. Only when the models are stressed with subtle data set differences does OPLS emerge as the best-performing model. OPLS maintained a high-prediction accuracy rate and a large area under the ROC curve while yielding loadings closest to the true loadings with limited group separations.



KEYWORDS: *metabolomics, multivariate, classification models, NMR*

■ INTRODUCTION

Metabolomics relies on the measurement of small-molecule concentration changes that result from perturbations in specific cellular processes.[1] In this regard, metabolomics aims to understand a system-wide response to an external stimulus, environmental stress, or a genetic adaptation. The metabolome is thus a better proxy for the state of a biological system since metabolites are the direct outcomes of all genomic, transcriptomic, and proteomic responses to these stimuli, stress, or genetic mutations.[2] Metabolomics has been experiencing exponential growth and, accordingly, has been applied to a variety of disciplines that includes food science and nutrition,[3] toxicology,[4] pharmacology,[5] and medicine.[6] Metabolomics is also playing an important role in drug discovery and precision medicine by identifying biomarkers associated with complex diseases such as atherosclerosis,[7] cancer,[8] diabetes,[9] and obesity.[10]

Metabolomics experiments are often complex and involve a large number of chemically diverse metabolites.[11] For example, the plant kingdom is estimated to contain over 200 000 metabolites,[12] and the human metabolome database currently contains over 114 000 entries.[13] NMR spectroscopy[14] and mass spectrometry (MS)[15] are routinely employed to characterize the metabolome for a wide range of sample types that includes cell lysates, tissues, organs, organisms, and complex biofluids (e.g., blood and urine).[16] Mass spectrometry also relies on the inclusion of liquid chromatography (LC),[17] gas chromatography (GC),[18] and/or capillary electrophoresis[19] techniques that further complicates the metabolomics data set. Given the diversity of the metabolome, these analytical platforms may generate tens of thousands of spectral

Received: April 8, 2019
Published: August 6, 2019

DOI: 10.1021/acs.jproteome.9b00227

features across numerous biological replicates.[20] Accordingly, metabolomics data is fundamentally multivariate in structure—multiple independent variables (i.e., chemical shifts, *m*/*z*, or retention times) for each biological replicate. As an example, for each NMR spectrum, the chemical shifts are independent variables, while resonance or peak intensities are the corresponding measurements. The total number of chemical shifts, depending on the spectral resolution, is typically on the order of 10^3−10^4 values.

Multivariate data analysis is often employed to detect relationships between the measured experimental variables in these large data sets to obtain insights regarding group membership.[21,22] The resulting models are then used to identify key variables that define these group memberships. These key variables, in turn, may be used as part of a predictive model or to understand the underlying data structure of the groups or the differences between the groups. Multivariate analysis is routinely used in lieu of univariate analysis because variables from the same metabolite, metabolites from the same metabolic pathway, or metabolites from coupled metabolic pathways tend to be highly correlated. Multivariate analysis takes advantage of the fact that some variables are correlated by simultaneously examining the entirety of the data set. Conversely, univariate analysis assumes that all of the observed variables are independent and simply relies on pairwise comparisons. Thus, multivariate data analysis has been an important component of a significant number of metabolomics studies (Figure 1) and is often employed to answer some fundamental questions about a metabolomics data set. This includes determining if biological samples from two or more groups actually differ. For example, does the chemical signature in urine differ between healthy controls and multiple sclerosis patients?[23] Given an observed difference between these groups, an additional goal is to identify and quantify the specific metabolites associated with each biological state. Again, for example, an NMR metabolomics study identified eight urinary metabolites (acetate, creatinine, hippurate, 3-hydroxybutyrate, 3-hydroxyisovalerae, malonate, oxaloacetate, and trimethylamine *N*-oxide) associated with multiple sclerosis that may be used as prospective biomarkers.[24]

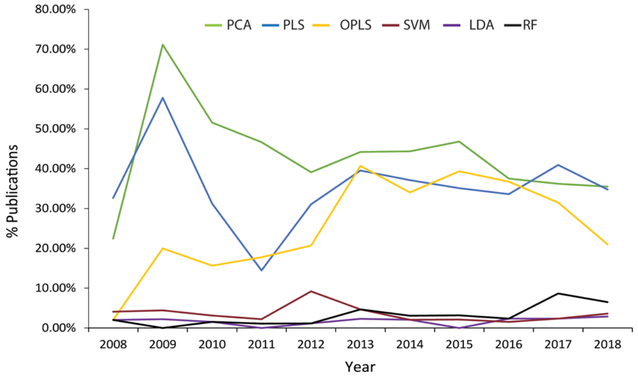


Figure 1. Plot of the percentage of NMR metabolomics publications using multivariate statistical analysis from 2008 to 2018 that included principal component analysis (PCA) (green), partial least-squares projection to latent structures (PLS) (blue), orthogonal projection to latent structure (OPLS) (yellow), support vector machine (SVM) (dark red), linear discriminant analysis (LDA) (purple), and/or random forests (RF) (black) data analysis.

Principal component analysis (PCA)[25] was the multivariate technique first introduced by Nicholson et al. (1999)[26] as a valuable approach to analyze NMR metabolomics data sets to identify group membership. While PCA is still a valuable and routinely used statistical tool for metabolomics,[27] PCA's inherent limitation was quickly recognized.[21] Specifically, PCA is an unsupervised technique that simply identifies the largest variance in the data set and defines group membership regardless of the source of the variance. Since a metabolomics study is typically designed with predefined group membership (i.e., healthy vs disease), the desired outcome of the multivariate model is to confirm the expected group membership and to identify the key metabolites correlated with these defined groups. Accordingly, metabolomics is now utilizing supervised classification models that include partial least-squares projection to latent structures (PLS),[28] linear discriminant analysis (LDA),[29] orthogonal projection to latent structure (OPLS),[30] the combination of PCA with LDA (PC-LDA),[31] and machine learning algorithms such as support vector machines (SVM)[32] and random forests (RF)[33] to identify the spectral features that define a group membership. PCA is still commonly used as a first-pass quality control method prior to employing a supervised classification approach.[27] In this regard, PCA provides an unbiased verification that a group variance actually exists in the data set, which may imply that a subsequent supervised model is also valid.[27] Figure 1 plots the change in multivariate classification models applied to NMR data sets over the last decade.

Initially, PLS dominated metabolomics studies, but OPLS quickly gained parity. Even though PLS and OPLS still dominate NMR metabolomics studies, machine learning models such as SVM and RF have started to garner some attention. Nevertheless, these variable trends in the application of multivariate data analysis raise some serious questions. Is the popularity of PLS and OPLS the result of applying the best model or is it due to other considerations? Are investigators simply applying a model based on their preference or familiarity, based on the model's availability in software, or simply because investigators are following literature precedence? Of course, the preferred answer would be for investigators to select the best model, but what defines the "best model"? Furthermore, the literature currently lacks any quantitative comparisons of common multivariate classification models to inform investigators of a best-choice model for metabolomics. To address these issues, we report herein a comprehensive comparison of multivariate classification models currently available in our open-source MVAPACK software package (http://bionmr.unl.edu/mvapack.php).[34] The multivariate classification models were evaluated using both simulated and experimental NMR data sets. Specifically, the data sets were designed to contain two groups with variable levels of group separation. Notably, the data sets were designed such that the variables defining group separation and the magnitude of within-group and between-group variances are all known quantities. Intuitively, a large difference between groups should be readily identifiable regardless of the model. However, model performance is expected to deteriorate as the within-group and between-group variances worsen, which usually occurs in real biological data sets, especially large clinical trials involving human patients. Accordingly, being able to select a multivariate classification model with a high-enough power to still identify subtle group differences in the presence

DOI: 10.1021/acs.jproteome.9b00227

Table 1. List of Metabolites Used To Construct the Simulated and Experimental Data Sets

simulation		NMR experiment	
acetic acid	L-cysteine	acetic acid	methanol
acetoacetic acid	L-cystine	anserine	*N*-acetylneuraminic
acetone	L-fucose	citric acid	sucrose
adipic acid	L-glutamine	creatinine	taurine
anserine	L-histidine	D-glucose	trigonelline
cis-aconitic acid	L-lysine	dimethylamine	trimethylamine
citric acid	L-serine	dimethylglycine	urea
creatine	L-threonine	D-mannitol	
creatinine	L-tyrosine	ethanolamine	
D-glucose	mandelic acid	gluconic acid	
dimethylamine	methanol	glycine	
dimethylglycine	methylguanidine	glycolic acid	
D-mannitol	methylsuccinic acid	guanidoacetic acid	
ethanolamine	*N*-acetylaspartic acid	hippuric acid	
formic acid	*N*-acetylneuraminic acid	imidazole	
gluconic acid	phenol	isobutyric acid	
glycine	pyroglutamic acid	isocitric acid	
glycolic acid	sucrose	D-lactic acid	
guanidoacetic acid	taurine	L-alanine	
hippuric acid	trigonelline	L-fucose	
imidazole	trimethylamine	L-glutamine	
isobutyric acid	urea	L-histidine	
isocitric acid	3-(3-hydroxyphenyl)-3-hydroxypropanoic acid (HPHPA)	L-lysine	
L-2-hydroxyglutaric acid	3-aminoisobutanoic acid	L-serine	
D-lactic acid		L-threonine	
L-alanine		L-tyrosine	

of high variance is essential to metabolomics. In this regard, five multivariate classification models currently employed by metabolomics investigators (PC-LDA, OPLS, PLS, RF, and SVM) and incorporated into our MVAPACK software package were evaluated based on their ability (i) to correctly predict group memberships of unseen samples, (ii) to maximize sensitivity and specificity, and (iii) to correctly identify true spectral features associated with group differences. Our findings suggest that the five classification models performed equally well with robust data sets, but OPLS performed best in the analysis of one-dimensional (1D) ^{1}H NMR metabolomics data sets with minimal group separation. While the performance of the five classification models was evaluated utilizing NMR data sets, the results are likely generalizable to other analytical data sources since the multivariate data structure would be similar to NMR.

■ MATERIALS AND METHODS

The performance of the five multivariate classification models was assessed using both simulated and experimental 1D ^{1}H NMR spectral data sets. All analyses were conducted using our MVAPACK software package.[34] All of the figures were generated using the R software package[35] and Excel.

Simulated 1D ^{1}H NMR Data Set

An artificial mixture consisting of 50 common urine metabolites (Table 1)[36] (e.g., creatinine, glycine, formic acid, isocitric acid, urea, etc.) was used to generate a simulated 1D ^{1}H spectrum with a total of 690 peaks (Figure S1). An NMR spectrum may be considered as a linear combination of peaks, which can be ideally generated using a Lorentzian line shape[37] (also known as the Cauchy distribution function). Peak locations (i.e., chemical shifts) and relative peak intensities for each metabolite in the artificial mixture were obtained from the human metabolome database (HMDB).[13] The simulated 1D ^{1}H NMR spectrum was approximated as

$$S(x) = \sum_{i=1}^{n} w_i{*}\frac{1}{1 + \left(\frac{x - l_i}{\gamma_i}\right)^2} \tag{1}$$

where l_i and γ_i are the peak location (ppm) and peak width (ppm) of the ith peak, respectively; w_i is a multiplier factor for the ith peak based on peak intensities from HMDB; n is the number of peaks; and x is all possible chemical shifts, which range from 0.9 to 9.2 ppm with an equal spacing of 0.001 ppm. γ_i was set to 0.002 ppm to match the typical peak shape in HMDB. Importantly, the multiplier factor w_i was determined such that relative peak intensities for each metabolite (as defined by the reference spectrum in the HMDB) were maintained even as additional peaks and metabolites were added to the simulated spectrum. Simply, the multiplier factor w_i is adjusted to accommodate partially overlapped peaks.

Each simulated NMR data set consisted of 100 spectra (N = 100), where each spectrum contained 8301 spectral features (p = 8301). Each data set was divided into two distinct groups (group 1 and group 2), where each group contained 50 spectra. Group 1 and group 2 were differentiated by changing the mean concentration of a single metabolite from the 50 metabolites that comprise the artificial mixture. Importantly, the mean concentration for the remaining 49 metabolites was kept constant between the two groups, where the mean concentration was set to one (arbitrary units). A total of nine (w1−w9) different NMR simulated data sets were created, which differed by the amount of group separation. The relative amount of group separation increased from w1 to w9. The group separation was defined by sequentially increasing the

Table 2. Description of the Nine Simulated Data Sets Comprising Three Scenarios

				mean concentration		sample concentration	
set	scenario	metabolite changed	# peaks	group 1	group 2	group 1	group 2
w1 (5%)	1	D-glucose	48	1	1 + 5% = 1.05	∼Unif(1 ± 0.1)	∼Unif(1.05 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w2 (7%)	1	D-glucose	48	1	1 + 7% = 1.07	∼Unif(1 ± 0.1)	∼Unif(1.07 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w3 (10%)	1	D-glucose	48	1	1 + 10% = 1.1	∼Unif(1 ± 0.1)	∼Unif(1.1 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w4 (15%)	1	D-glucose	48	1	1 + 15% = 2	∼Unif(1 ± 0.1)	∼Unif(1.15 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w5 (20%)	1	D-glucose	48	1	1 + 20% = 1.2	∼Unif(1 ± 0.1)	∼Unif(1.2 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w6 (30%)	1	D-glucose	48	1	1 + 30% = 1.3	∼Unif(1 ± 0.1)	∼Unif(1.3 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w7 (40%)	1	D-glucose	48	1	1 + 40% = 1.4	∼Unif(1 ± 0.1)	∼Unif(1.4 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w8 (50%)	1	D-glucose	48	1	1 + 50% = 1.5	∼Unif(1 ± 0.1)	∼Unif(1.5 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				
w9 (100%)	1	D-glucose	48	1	1 + 100% = 2	∼Unif(1 ± 0.1)	∼Unif(2 ± 0.15)
	2	isocitric acid	15				
	3	methanol	1				

mean concentration of a single metabolite in group 2 relative to group 1. The mean concentration of a single metabolite in group 2 was increased by 5, 7, 10, 15, 20, 30, 40, 50, and 100%, respectively. Importantly, there were three different scenarios (1−3) within each of these nine simulated data sets. The three scenarios were differentiated by the identity of the metabolite that was subjected to the sequential increase in the mean concentration. In scenario 1, the mean concentration of D-glucose, which has 48 NMR peaks, was changed between groups 1 and 2. The mean concentration of isocitric acid (15 NMR peaks) was changed in scenario 2, and methanol (1 NMR peak) was changed in scenario 3. Independent noise from a Gaussian distribution (mean = 0, standard deviation (SD) = 5% of group mean spectrum) was added to each spectrum to represent systematic variability. Biological variability was incorporated into each data set by randomly varying each of the metabolite concentrations about its mean within a defined range (±0.1 for group 1, ±0.15 for group 2). In summary, 27 different simulated NMR data sets were created, where each data set contained 100 simulated spectra for a total of 2700 simulated 1D ^{1}H NMR spectra. Table 2 summarizes the composition of the entire simulated NMR data set.

Experimental 1D ^{1}H NMR Data Set

Saturated solutions were prepared for the 33 common urine metabolites listed in Table 1. Each individual metabolite was dissolved to saturation in 1 mL of NANOPure water (Barnstead, Dubuque, IA). NMR samples were prepared by first diluting 10 μL of the stock metabolite solution with ddH_2O to a final volume of 30 μL. The diluted metabolite solution was then added to 570 μL of a 50 mM phosphate buffer in D_2O at pH 7.2 (uncorrected) for a final volume of 600 μL. The phosphate buffer solution also contained 500 μM of 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid sodium salt (98% D) (TMSP-d_4) as an internal chemical shift and concentration standard. The samples were transferred to a 5 mm NMR tube for data collection. A 1D ^{1}H NMR spectrum was collected for each of the metabolites. The NMR spectra were collected at 298K with 64 scans and 4 dummy scans and a 2 s relaxation delay. The spectra were collected with a spectral width of 11 160 Hz, 32 K data points, and excitation sculpting[38] to suppress the solvent resonance and maintain a flat baseline. NMR spectra were processed using our MVAPACK software package.[34] The water signal between 4.65 and 4.9 ppm was removed prior to creating an artificial urine mixture (Figure S2) to generate binary classification data sets (e.g., control vs treated) with various amounts of group separation following the same scenarios described above for the simulated 1D ^{1}H NMR data set (Table 2).

Preprocessing of NMR Data Sets

To emphasize the difference between the five multivariate classification models, a fixed and minimalistic data preprocessing protocol was employed. The 1D ^{1}H NMR spectrum for each metabolite was normalized to its most intense peak prior to adding the spectra together to create the NMR spectrum for each mixture. The input data for the classification algorithms was then presented as a matrix of rows and columns corresponding to biological replicate samples (i.e., replicate 1D ^{1}H NMR spectra) and spectral features (i.e., the individual spectral data points). Each data set was then scaled

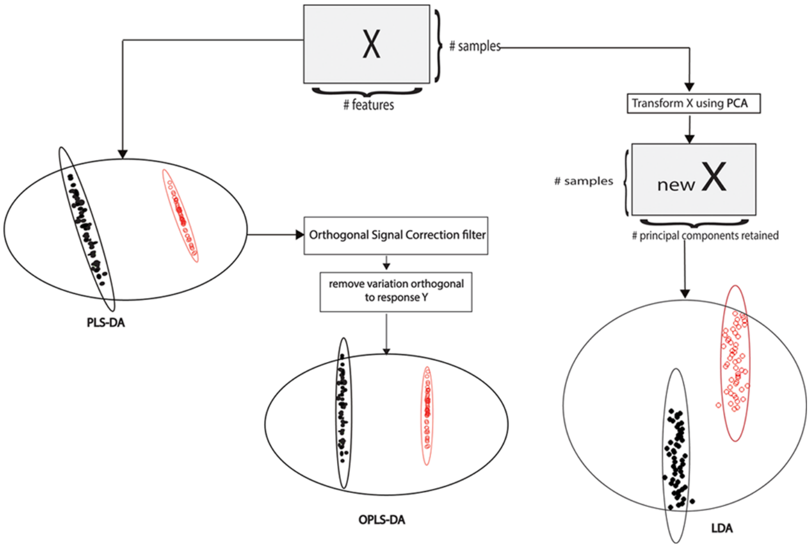

X
# samples
# features
Transform X using PCA
# samples
new X
# principal components retained
Orthogonal Signal Correction filter
remove variation orthogonal to response Y
PLS-DA
OPLS-DA
LDA

Figure 2. Classification process using PLS, OPLS, and PC-LDA models. The process starts with a data matrix X with rows and columns representing biological samples and predictor features, respectively. On the left, the data matrix X is submitted to PLS, which results in a two-dimensional scores plot with separation between groups 1 and 2. In the middle, data matrix X is submitted to OPLS, which is similar to PLS, but with an additional orthogonal signal correction (OSC) filter to remove confounding variation that is not explained by response (i.e., group membership) variables. The OSC filter rotates the resulting scores plot. On the right, data matrix X is first submitted to PCA to transform the data into a new matrix (new X) with rows and columns representing biological samples and retained principal components, respectively. The new X is then submitted to LDA to obtain a scores plot. Red and black points represent samples in group 1 and group 2, respectively. Small ellipses are class-specific 95% confidence ellipses, while the large ellipses are the overall 95% confidence ellipses.

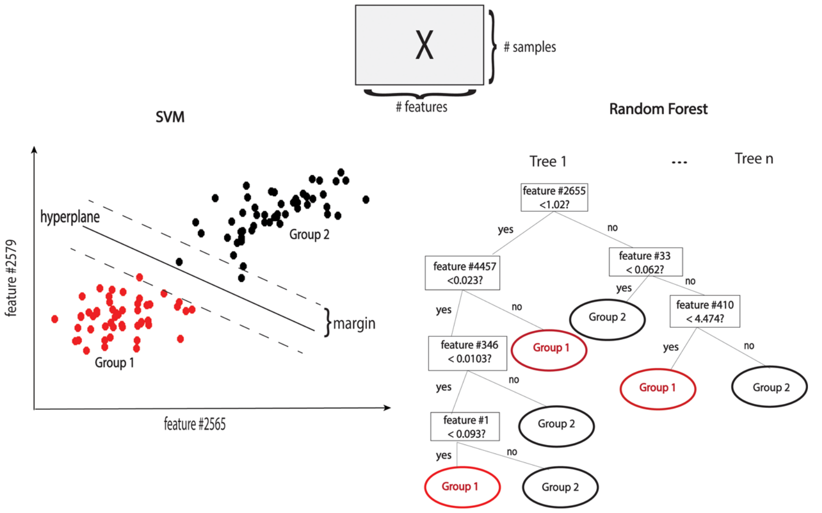


Figure 3. Classification process using SVM and RF models. The process starts with a data matrix X with rows and columns representing biological samples and predictor features, respectively. Feature numbers specified in the plot are presented as indexes along the column of data matrix X. On the left, data matrix X is submitted to SVM, which finds an optimal hyperplane with a maximized margin to separate the biological samples into two groups. On the right, data matrix X is submitted to random forests, which combines *n* decision trees. Each tree is a chain of binary splits, where each split evaluates one feature at a specific value. This is demonstrated by the series of decision boxes in the tree. Each biological sample is passed through each consecutive decision conditions until assigned to group 1 or group 2.

columnwise by its own mean and variance. There was no missing data in the data sets.

Classification Algorithms

Partial Least-Squares Projection to Latent Structures (PLS). The PLS algorithm is an extension of partial least-square regression to the classification problem, which was illustrated to chemometrics pattern recognition by Dunn and Wold in 1990.[39]PLS finds latent variables that maximize the correlation between predictor variables (X) and the categorical response variable (Y) while reducing data dimensions. PLS has been shown to be robust in handling highly correlated predictor variables, which are common outputs in metabolomics experiments.(Figures 2 and 3)

Orthogonal Projection to Latent Structure (OPLS). OPLS is similar to PLS, in that the algorithm maximizes a relationship between the predictor and categorical response variables through latent variables. However, OPLS takes advantages of an orthogonal signal correction (OSC) filter[40] to remove variations in the predictor variables that are not explained by the response. As a result, the separation between observations in the latent space (i.e., scores space) is improved.

Support Vector Machines (SVM). Support vector machine is a machine learning tool developed by Vapnik.[41] The SVM algorithm finds an optimal separating hyperplane with a maximum distance to the training observations, which is called margin. In other words, when classes are overlapped, SVM is constructed by minimizing the cost of the training points that are on the wrong side of the classification boundary. SVM can also be extended to nonlinear boundaries by utilizing kernel functions to map the training observations to a higher-dimensional space.[42]

Random Forests (RF). Random forests is an ensemble learning method that combines multiple decision trees to predict group membership by majority votes.[33] Each decision tree in the forest is trained on a random subset of the observations and then tested on the remaining observations to avoid overfitting. The forest is tuned to the number of trees and to the number of random variables used at each split.[43]

Combination of Principal Component Analysis and Linear Discriminant Analysis (PC-LDA). Linear discriminant analysis (LDA) is a classical statistical algorithm for obtaining a classification rule by maximizing the variation between group memberships relative to in-group variability.[44] In other words, the LDA algorithm finds a combination of predictor variables that maximizes the distance between the centers of two groups while minimizing the variation within each group. However, the LDA algorithm involves inverting a covariance matrix, which requires a large number of samples (N) relative to the number of variables (p). As a result, LDA cannot be directly used on data sets that have more variables than samples ($p > N$) and/or have highly correlated variables. One popular solution is to utilize a dimension reduction technique such as principal component analysis (PCA). PCA transforms the original data set to uncorrelated predictors (i.e., principal components) prior to applying LDA.[43] In this regard, the number of variables is greatly reduced while still retaining important data information.

Evaluation Criteria

Machine learning models require the specification of hyperparameters, such as cost values, kernel functions, and the number of trees to train a data set. SVM models were initiated with a vector of 20 cost values ranging from e^1 to e^{10}, where the ratio of two consecutive cost values was fixed at $e^{0.5}$. The SVM models used a linear, radial basis, or a third-degree polynomial kernel function. The resulting kernel function and cost value were chosen such that the misclassification rate was minimized by five iterations of fourfold cross validation. Specifically, 75% of the data set was used to train the SVM model and 25% of the data set was used for testing the SVM model. In a similar manner, the RF model was trained using 50, 100, 150, or 200 number of trees. The number of trees selected for the RF model yielded the smallest misclassification rate in cross validation. A data set of 70% was used to train the RF model and 30% of the data set was used for testing the RF model.

Classification Accuracy Rate. Each data set was randomly partitioned into four equal folds, where each fold contained 25 observations. Each model was trained on three folds ($N = 75$), and the model was then used to predict the remaining fold ($N = 25$). The process was repeated until each fold was used as the prediction fold. The entire process (i.e., partitioning, training, and prediction) was repeated three times to minimize any unintended basis in the partitioning of the data set into the four folds. The classification accuracy rate (AR_{ij}) for the ith partition ($i = 1, 2, 3, 4$) and the jth iteration ($j = 1, 2, 3$) was calculated as

$$\mathrm{AR}_{ij} = \frac{c_{ij}}{n_{ij}} \tag{2}$$

where c_{ij} is the number of correctly predicted observations and n_{ij} is the total number of observations in the ith partition and the jth iteration. AR_{ij} values range between 0 and 1, where an AR_{ij} value of 1 would indicate a perfect model performance.

Area Under Receiver Operating Characteristic (AUROC) Curve. Each data set was randomly partitioned into four equal folds, where each fold contained 25 observations. Each model was trained on the three folds ($N = 75$), and the model was then used to predict the remaining fold ($N = 25$). At the prediction step, the posterior probability for each new observation was calculated instead of using the group membership produced by the algorithm. Each posterior probability was then used as a threshold to assign group membership and to obtain the corresponding true positives and false positives. A ROC curve and AUROC were obtained for each partition and iteration, and a mean AUROC and standard deviation were then calculated. AUROC values range between 0.5 and 1, where an AUROC value of 1 would indicate a perfect model performance.

Root-Mean-Squared Estimates of Model Loadings. A reference set with "perfect" or maximal group separation under each scenario listed in Table 2 was generated. The loadings consisted of the contributions from each spectral feature to each classification model. Each data set was then compared to the true loadings from the reference set with perfect separation. Figure 4 illustrates loadings obtained from the OPLS model at

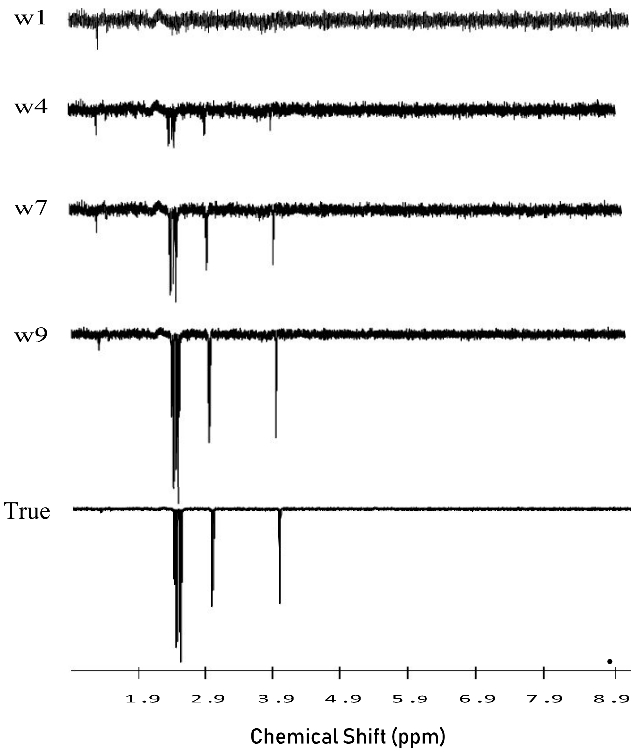


Figure 4. Plots of the loadings generated from the simulated data set OPLS model for different group separations w1, w4, w7, and w9 as defined in Table 2. The OPLS model loadings are compared against the true loadings.

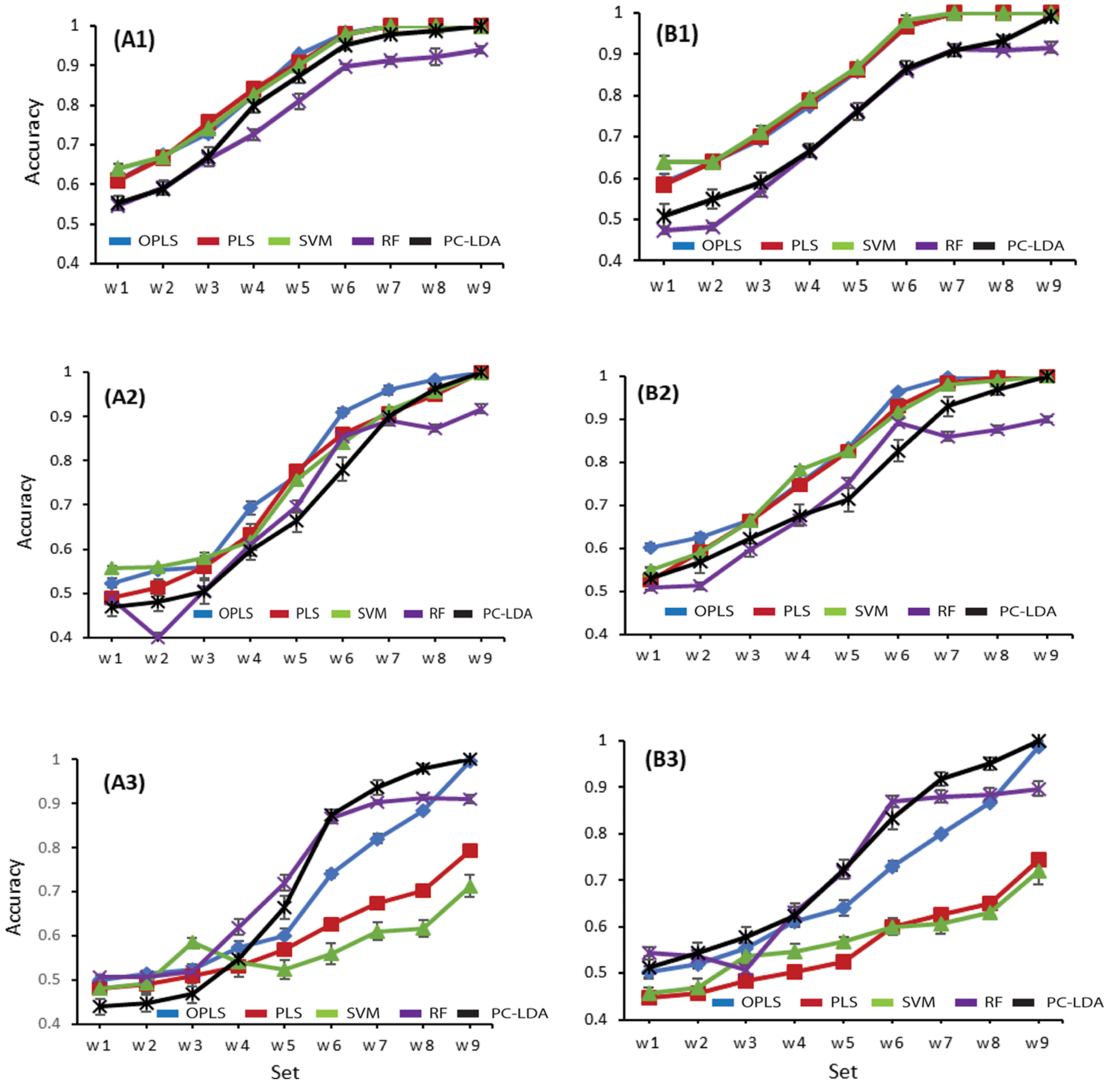


Figure 5. Classification accuracy rates for the five models: OPLS (blue), PLS (red), SVM (green), RF (purple), and PC-LDA (black) as calculated from the simulated (A) and experimental (B) data sets. The classification accuracy rates are plotted as a function of group separation (w1−w9), as defined in Table 2. The results from each of the three scenarios corresponding to varying (A1, B1) D-glucose, (A2, B2) isocitric acid, and (A3, B3) methanol were plotted separately. The classification accuracy rates are plotted as a mean ± SD.

four different group separations w1, w4, w7, and w9, respectively. Root-mean-squared estimates (RMSE_{ik}) were calculated between the loadings of the ith set resulting from the kth model ($\mathbf{x}_{ik}$) and the associated true loadings of the reference set ($\mathbf{t}_k$) using the following formula

$$\text{RMSE}_{ik} = \sqrt{\sum_{j=1}^{p} (\mathbf{x}_{ikj} - \mathbf{t}_{kj})^2} \tag{3}$$

where $\mathbf{x}_{ikj}$ is the jth element of the vector $\mathbf{x}_{ik}$, $\mathbf{t}_{kj}$ is the jth element of the vector $\mathbf{t}_k$, p is the number of spectral features, and k is the index of each model. RMSE_{ik} values were normalized by dividing each element by the maximum element. RMSE_{ik} values range between 0 and 1, where an RMSE_{ik} value of 0 would indicate a perfect model performance.

Loadings are not generated from an RF model. Instead, variable importance was used as a replacement for loadings in the RMSE criterion. RF variable importance was measured by the improvement in splitting observations accumulated over every node of all trees in the forest. The improvement measurement was described by the mean decrease in the Gini index.[29] In particular, a variable importance vector was obtained for each data set and the reference set, which was then normalized by its maximum value prior to computing an RMSE value according to eq 3. In this regard, an RMSE criterion can be directly compared between the five classification models.

■ RESULTS AND DISCUSSIONS

Evaluation Criteria

PC-LDA, OPLS, PLS, RF, and SVM models were generated for each of the simulated and experimental 1D ^{1}H NMR data sets summarized in Table 2. The relative performance of the five multivariate classification models was assessed by calculating a classification accuracy rate (AR_{ij}) (Figure 5), an area under receiver operating characteristic curve (AUROC) (Figure 6), and a root-mean-squared estimate of model loadings (RMSE_{ik}) (Figure 7) for each model.

DOI: 10.1021/acs.jproteome.9b00227

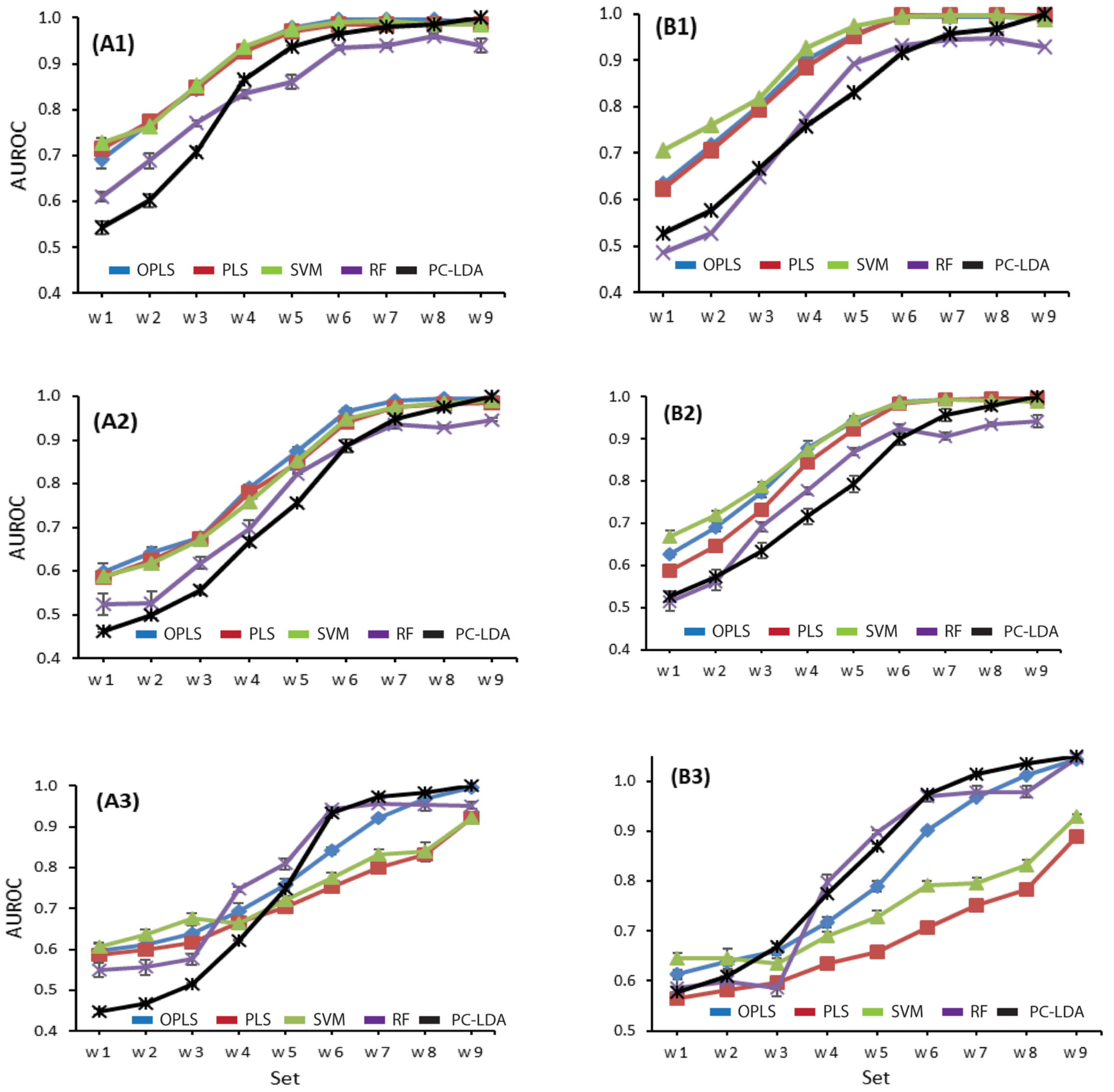


Figure 6. Area under a ROC curve for the five models: OPLS (blue), PLS (red), SVM (green), RF (purple), and PC-LDA (black) as calculated from the simulated (A) and experimental (B) data sets. The AUROCs are plotted as a function of group separation (w1−w9), as defined in Table 2. The results from each of the three scenarios corresponding to varying (A1, B1) D-glucose, (A2, B2) isocitric acid, and (A3, B3) methanol were plotted separately. AUROCs are plotted as a mean ± SD.

A common method to assess the performance of a classification model is to measure the overall classification accuracy rate (eq 2), which measures the correlation between the observed and predicted group memberships. Accordingly, a perfect model performance would yield a classification accuracy rate of 1. However, the overall classification accuracy rate does not account for differences in misclassification costs (e.g., false positives, false negatives) and the imbalance in natural group frequencies. Thus, an overall classification accuracy rate is not sufficient to measure model performance.[45] Additional statistical parameters need to be included, such as sensitivity and specificity. Sensitivity (i.e., true-positive rate) measures the number of correctly predicted positives out of all of the true positives. Similarly, specificity (i.e., true-negative rate) measures the number of correctly predicted negatives out of all of the true negatives. A classical evaluation framework that combines both the true-positive rate and the false-positive rate is a ROC curve.[46] A ROC curve plots pairs of true-positive and false-positive rates at different threshold values. An ideal classifier would completely separate the two classes with 100% sensitivity and specificity and yield an area under the ROC curve (AUROC) of 1. Conversely, an ineffective classifier would result in a ROC curve along the diagonal with an area under the curve close to 0.5.[43] Thus, a larger AUROC implies a better classification model and is a useful metric to compare multiple predictive models in combination with the classification accuracy rate.

In addition to predictive ability, the correct identification of the underlying biological factors that give rise to the group differences is a critical consideration of model performance. How well does the model identify the true discriminatory features? To address this issue, a reference set with perfect or maximal group separation under each scenario listed in Table 2 was generated to assess a model's ability to regenerate these true discriminatory features. In this regard, the ability of a model to reproduce the true discriminatory features for a given scenario was assessed by comparing the loadings from each model to the true loadings from the reference set. A high-performing model should perfectly reproduce the true loadings and yield an RMSE$_{ik}$ value (eq 3) of 0. Thus, to truly assess the

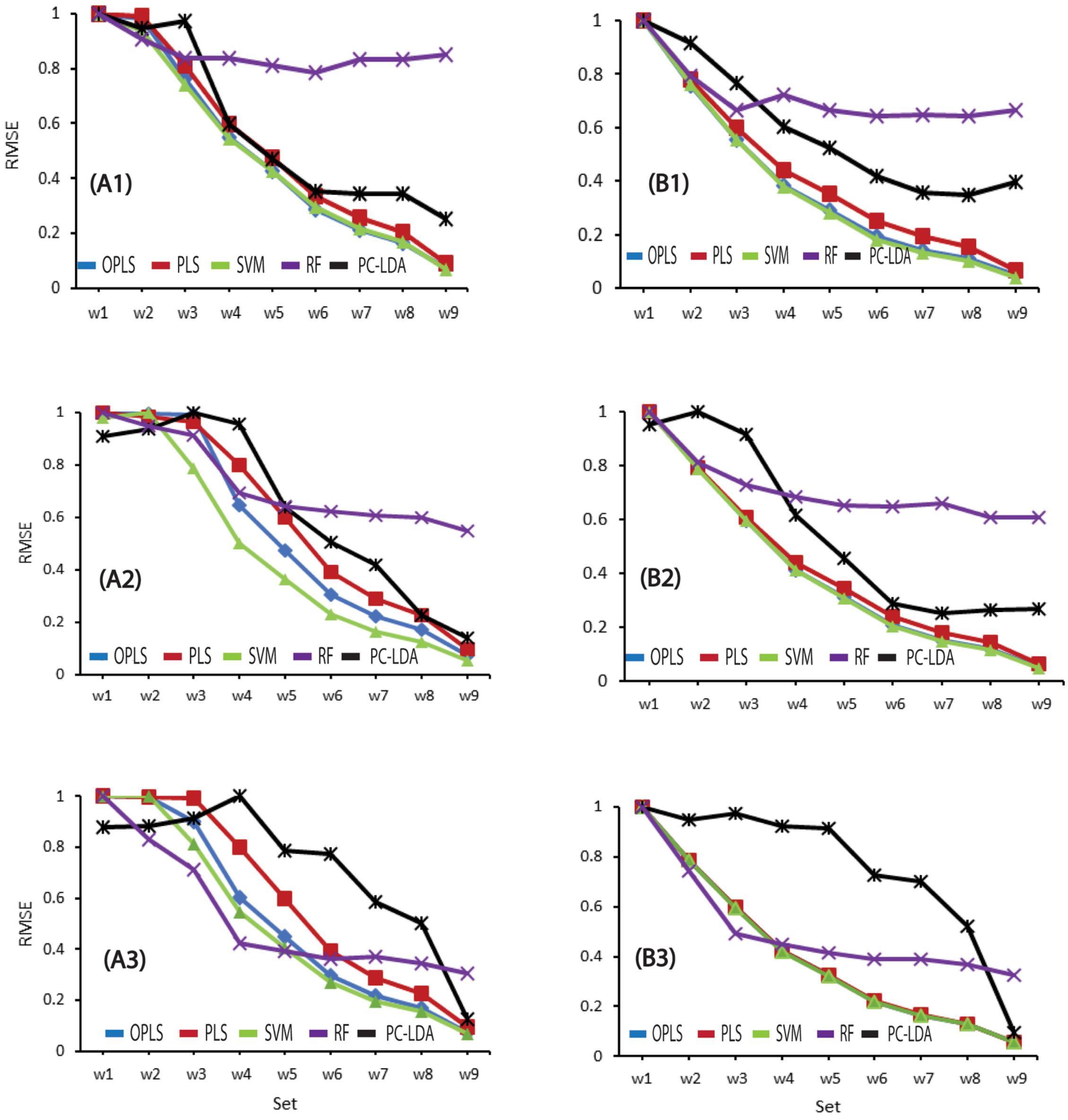


Figure 7. Root-mean-squared estimates (RMSE_{ik}) between each model's loadings: OPLS (blue), PLS (red), SVM (green), RF (purple), and PC-LDA (black) and the associated true loadings of the reference set as calculated from the simulated (A) and experimental (B) data sets. The RMSE_{ik} values are plotted as a function of group separation (w1−w9), as defined in Table 2. The results from each of the three scenarios corresponding to varying (A1, B1) D-glucose, (A2, B2) isocitric acid, and (A3, B3) methanol were plotted separately. RMSE_{ik} are plotted as a mean ± SD.

overall performance of the multivariate classification models, it is necessary to simultaneously consider these three evaluation criteria. The best- or highest-performing multivariate classification model should outperform the other models in all three categories. Conversely, mixed or similar evaluation criteria outcomes between two or more models would suggest similar model performance.

Simulated versus Experimental 1D ^{1}H NMR Data Sets

The simulated 1D ^{1}H NMR data set was designed to mimic an NMR spectrum obtained for a human urine sample. In this regard, chemical shifts and relative peak heights for 50 metabolites routinely observed in human urine samples were used to construct the simulated NMR spectra.[36] Importantly, the simulated NMR data set enabled the direct control and construction of the data structure. Specifically, the within-group and between-ground variances were known quantities. Similarly, the spectral features that defined group separation were also precisely defined. Accordingly, the performance of the five classification models could be accurately assessed since the outcomes were a known quantity. It would not be possible to achieve this same level of certainty with real biological samples. Nevertheless, there is always a concern that simulated data may oversimplify the system, may introduce unintended bias, or may miss important variables. Thus, to partially address these concerns, a second NMR data set was created from experimental 1D ^{1}H NMR spectra.

Of the 50 metabolites, 33 used in the simulated 1D ^{1}H NMR data set were commercially available and were used to collect an experimental 1D ^{1}H NMR spectrum on a 700 MHz spectrometer equipped with a cryoprobe. An NMR spectrum was collected for each individual metabolite. These individual experimental NMR spectra were then used to construct the same data set, as outlined in Table 2. Simply, the experimental 1D ^{1}H NMR spectrum for each individual metabolite was combined and scaled accordingly to achieve the desired within-group and between-group variances obtained with the fully simulated data set. While the experimental NMR data set is still properly defined as synthetic, it does incorporate various experimental features that are difficult or impossible to simulate, such as variations in instrument performance, sample preparation, and spectral noise. Accordingly, the primary intent of the experimental NMR data set was to validate the results from the simulated NMR data set and to confirm the absence of any unintended bias. As expected, the model performance against both the simulated and experimental NMR data sets was very similar, if not nearly identical. This is evident by comparing panels A1−A3 to B1−B3 in Figures 5−7. Thus, the assessment of model performance is reproducible and not data-set-dependent.

Performance of the Multivariate Classification Models

The most notable outcome from the analysis of the five commonly used metabolomics classification models is their overall equivalent performance with high-quality data sets: low noise, small within-group variance, and large between-group variance. All of the models routinely yielded perfect group membership and identified the “true” loadings when a clear group separation existed in the data set. In fact, to observe any difference in model performance, it was necessary to stress the situation beyond what is typically expected for an experimental metabolomics data set. For example, varying multiple metabolite concentrations as outlined in Table 2 (data not shown) between the two test groups would always yield a perfect outcome regardless of the model. Even limiting the test data sets to only one varying metabolite could still easily yield perfect model performance (Figures 5−7) unless the metabolite concentration difference or the number of NMR resonances was minimized. In fact, the largest variance in model performance was achieved when the variant metabolite was methanol with a single NMR peak (scenario 3).

As expected, there were common trends in model performance against both the simulated and experimental NMR data sets. There were also common trends across the three scenarios and as a function of group separation. For example, the classification accuracy and AUROC increased for all models as the group separation increased from data sets w1 to w9. Similarly, RMSE$_{ik}$ decreased as the group separation increased from data sets w1 to w9. Again, this observation highlights that the primary factor that determines model performance is the intrinsic size of group separation, not the model type. Differences in model performance only become apparent when the evaluation criteria were compared across the three scenarios.

The only difference between the three scenarios was the number of NMR peaks that were varied between the two groups, which decreased from 48 peaks for D-glucose in scenario 1 to one methanol peak in scenario 3. A greater difference in model performance was observed as the number of peaks decreased, where the greatest difference occurred when only a single peak or spectral feature was varied between the two groups. Again, this highlights the fact that the five models, in general, all performed equally well. It was only under an extreme scenario, the small variance of a single peak, that model performance deviated. As an illustration, OPLS, PLS, and SVM performed modestly better (p-value <0.03) in classification accuracy (>6−12%) than RF and PC-LDA in scenario 1 (Figure 5 and Table S1A,B). However, the relative performance was nearly reversed in scenario 3, where RF, PC-LDA, and OPLS exhibited a significantly higher (p-value <0.05) classification accuracy (>2−33%) than SVM and PLS. A similar trend was observed for AUROC (Figure 6 and Table S2A,B). OPLS, PLS, and SVM had significantly higher AUROC (p-value <0.03) than RF and PC-LDA (>3−18%) for scenario 1, while the differences decreased in scenario 2. However, a significant divergence in the ROC curves (Figures S3 and S4) occurred between w6 and w9 for scenario 3. At the maximum group separation (w9), only the ROC curves for the OPLS and PC-LDA models yielded an AUC close to 1, while both SVM and PLS had an AUC between 0.84 and 0.92. In total, the five classification models performed equally well and as expected in terms of classification accuracy, sensitivity, and specificity. The model performance improved proportionally to the increase in group separation. Only when the spectral features defining group separation were limited to a single peak did OPLS and PC-LDA demonstrate improved performance over the other models.

In addition to correctly predicting group membership with a high level of sensitivity and specificity, a successful NMR metabolomics study also requires the proper identification of the classifying features. An RMSE$_{ik}$ was calculated between each model’s loadings and the true loadings to assess how well each model successfully identified the true discriminatory features. In other words, how well did the model perform in correctly identifying the correct metabolite changes. As expected, the deviation between each model’s loadings and the true loadings decreased as group separation increased (Figure 7 and Table S3A,B). However, the rate at which RMSE$_{ik}$ decreased varied significantly when comparing OPLS and SVM versus the other three models (p-values <0.03). OPLS and SVM performed the best across the entire data set and exhibited the fastest reduction in RMSE$_{ik}$. In fact, OPLS and SVM performed nearly identically except for simulated scenario 2 in which SVM performed modestly better.

Similar to classification accuracy and AUROC, the RMSE$_{ik}$ change rate was scenario-dependent. The rate of change in RMSE$_{ik}$ was fastest in scenario 1 and slowest in scenario 3. A notable exception was the RF model where the trend was reversed. In scenario 1, the RMSE$_{ik}$ for RF remained above 0.6, while the other models had an RMSE$_{ik}$ close to 0.1. However, the RMSE$_{ik}$ gap between RF and the other models was reduced as the number of varying peaks associated with group separation decreased. In fact, in scenario 3, RF had the smallest RMSE$_{ik}$ value for minimal group separations (w1−w3) but was outperformed by OPLS, SVM, and PLS at larger group separations (w4−w9). Thus, RF performed poorly for scenarios 1 and 2, but PC-LDA performed poorly in scenario 3. Conversely, OPLS and SVM performed best overall in terms of correctly identifying the true discriminatory features.

Choosing a Multivariate Classification Model

Despite the observation that all five classification models performed equally well with robust data sets containing clear group separations, the algorithms are not equivalent (Figures 2 and 3).[21] Accordingly, classification models are not interchangeable. Instead, each model provides a distinct interpretation or view of the data set. For example, PCA will identify the largest variance in the data set, but the largest variance may not be particularly relevant to the intended goal of defining the group differences, like differentiating between healthy controls and patients based on a disease pathology.

Instead, PCA may define group membership from some combination of confounding factors like diet, gender, race, etc. Thus, PCA is a valuable tool to verify that a group variance does exist, but it may provide misleading group-identifiable features. The same limitations apply to PC-LDA since it also relies on PCA. PC-LDA does have the advantage of simplifying the visualization of a model defined by multiple principal components. A score plot from most PCA models can be easily displayed using two or three principal components, but even a three-dimensional PCA score plot can be challenging to view. The situation becomes challenging, if intractable, when a correct PCA model requires more than three principal components. Simply, LDA allows for the projection of a multiple component model back down into two-dimensions while maintaining the group separation achieved in multiple PC space (Figure 2). Notably, PCA is an unsupervised technique, but PC-LDA is supervised. So, PCA provides an unbiased view of the data set and may be valuable for validating a supervised PLS or OPLS model.[27]

Supervised techniques, like PLS and OPLS, are useful for identifying the spectral features that define group separation and are usually employed after a PCA model demonstrates a clear group difference. However, PLS and OPLS models can be misleading and yield erroneous biological interpretations. This arises because PLS and OPLS will almost always produce the requested group separation, even for random noise.[47] Thus, PLS and OPLS models always require validation before proceeding with any model interpretation. Model validation is usually achieved with CV-ANOVA[48] or cross-permutation testing.[49] It is important to note that the R^2 and Q^2 values commonly reported for PLS/OPLS models do not provide a measure of model validation. R^2 and Q^2 are only indicators of model fit or a measure of model consistency, respectively. An important distinction between OPLS and PLS is how the algorithms handle cofounding factors (Figure 2). For OPLS, the spectral features related to the group differences are placed into the predictive component (x-axis). All of the spectral features associated with confounding factors are placed into the orthogonal component (y-axis). Conversely, PLS intermingles both group-defining features and confounding features leading to an apparent tilt or rotation in the scores plot (Figure 2). Thus, PLS may erroneously associate confounding spectral features with a valid group separation, and for this reason, PLS should be avoided for metabolomics analysis.

Conceptually, PCA, PLS, and OPLS fit the multidimensional data with a series of linear vectors or principal components ($\overrightarrow{\mathrm{PC}}_1$, $\overrightarrow{\mathrm{PC}}_2$, $\overrightarrow{\mathrm{PC}}_3$, etc.), where each vector is orthogonal to the previous one. The first vector captures the largest variance between the two groups, and each subsequent vector describes less of the data variance. In this regard, the data can be overfit by including a large number of principal components. SVM differs from PCA, PLS, and OPLS by fitting the data to a single hyperplane such that the hyperplane maximizes the separation between two groups (Figure 3).[41] The major advantage of SVM over PCA, PLS, and OPLS is the ability to fit the data with either a linear or nonlinear algorithm. SVM can use a variety of different kernel functions that include linear, polynomial, Gaussian, sigmoid, spline, etc. to fit the data. Like PLS and OPLS, SVM is prone to overfitting and requires validation.[50] In fact, the overfitting problem is correlated with the type of kernel function used and the value of the cost function (the margin width of the hyperplane, Figure 3). So, SVM is valuable if the data fits a nonlinear model, but it may be challenging to identify a proper kernel and cost function.

RF is probably the most unique classification model of the five algorithms investigated and provides a distinct approach to the analysis of the data set (Figure 3).[33,43] As the name implies, RF creates a decision tree by conducting a chain of binary splits based on the value (magnitude) of a specific spectral feature. Each decision is applied across the entire set of biological replicates or NMR spectra. As illustrated in Figure 3, if a spectrum has feature number 2655 (first decision box) with a value less than 1.02, then the spectrum proceeds to the next decision box based on feature number 4457. Otherwise, it follows an alternative decision path based on feature number 33. At the decision box for feature 4457, if the magnitude of the feature is greater than 0.023, then the spectrum is assigned to group 1. Otherwise, the process continues to the next decision box until the spectrum is classified into either group 1 or group 2. The process is highly biased by the starting point or defined path. In the decision tree in Figure 3, group membership is determined by the order of comparison to spectral feature numbers 4457 and 346. Thus, numerous trees need to be constructed from random starting points and different decision paths. RF has a number of unique advantages relative to the four other classification models. RF is better suited for data sets that contain multiple groups, where PLS, OPLS, and SVM are really designed for a comparison between two groups. Also, RF works well with a mixture of data sources and data types, which again, is problematic for the other classification methods. RF results are also easier to interpret; the output is a classification tree. Of course, RF needs to be validated like PLS, OPLS, and SVM, since it tends to overfit on training sets. Also, knowing how many trees to generate and obtaining a consensus tree may be a challenge. Thus, RF is a valuable choice for a data set that comprises multiple groups and various data types.

While PCA, PC-LDA, PLS, OPLS, SVM, and RF may perform equally well on a given data set, they are fundamentally distinct algorithms with unique advantages and limitations. Therefore, it is important to stress that these other factors need to be considered when choosing a particular multivariate classification model to analyze a specific data set. What are the important features or characteristics of the data set, what are the goals of the study, and what are the desired outcomes?

■ CONCLUSIONS

Five classification models, OPLS, PC-LDA, PLS, RF, and SVM, are routinely used by metabolomics investigators, but it is unclear which, if any, of these models are the best choice for the analysis of metabolomics data sets. Toward this end, the five models were evaluated based on classification accuracy rate, sensitivity and specificity, and the correct identification of the true discriminant features. The performance was assessed using both a simulated and experimental 1D ^{1}H NMR metabolomics data set in which the group separation and discriminant features were varied. Essentially equal results were obtained with both the simulated and experimental data sets demonstrating the robustness of the analysis and verifying that the results are data-set-independent.

All five models were observed to perform equally well when the data set contained clear group separation. The models perfectly predicted group membership and correctly identified the true discriminant features. Importantly, model performance was strongly correlated with group separation—as group separation increased, model performance increased. Thus, the choice of a model is irrelevant for a robust data set. So, any of the five classification models would be an acceptable and equivalent choice for the majority of metabolomics studies.

A difference in model performance was only observed with a relatively extreme data set structure. Specifically, model performance differed only with a data set consisting of a single metabolite change comprising a minimal number of NMR peaks with a limited peak intensity variance. In fact, the classification accuracy rate, AUROC, and RMSE all improved proportionally with an increasing number (from 1 to 48) of variable NMR peaks. One concerning inconsistency occurred with RF and PC-LDA. RF and PC-LDA outperformed the other models in classification accuracy rate and AUROC when a single feature (i.e., methanol) differentiated the two groups. However, RF and PC-LDA performed poorly in correctly identifying the true discriminant features (high RMSE). In essence, the correct group membership was identified using the wrong features. Overall, our findings indicate that OPLS was the best-performing model when considering overall performance with both the robust high-quality data sets and the extreme data set with minimal between-group variance. OPLS identified the true discriminant features while maintaining a reasonably high classification accuracy rate and a high AUROC. While all of the models performed well and are an acceptable choice with robust data, OPLS maintained its performance with minimal group separation. Thus, we would recommend OPLS as a preferred routine choice for the analysis of metabolomics data sets.

While the performance of the five classification models was evaluated using simulated and experimental 1D ^{1}H NMR data sets, the analysis is applicable to any other analytical source. LC-MS, GC-MS, or Fourier transform infrared (FTIR) multivariate data structure would be similar to NMR. So, the model performance and results are likely generalizable to any analytical source. To clarify, the input to the classification algorithms is a simple matrix of rows (biological replicates) and columns (variables). Specifically, the columns are abundance measurements across all observed variables. For NMR, the variables are defined as chemical shifts. For LC-MS or GC-MS, the variables would be defined as paired m/z and retention times, and for FTIR, the variables would be defined as wavenumbers. However, from the perspective of the classification algorithm, the label assigned to the variable is irrelevant and not part of the calculation. More importantly, the resulting LC-MS, GC-MS, or FTIR covariance matrix would be similar to that of the NMR data since measurements coming from the same metabolite, or from metabolites in the same metabolic pathway, or from metabolites in coupled metabolic pathways would still be highly correlated.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00227.

Pairwise p-values between classification models based on classification accuracy (Table S1); pairwise p-values between classification models based on AUROC (Table S2); pairwise p-values between classification models based on RMSE (Table S3); representative 1D ^{1}H NMR spectra for simulated data set (Figure S1); representative 1D ^{1}H NMR spectra for experimental data set (Figure S2); ROC curves from OPLS or PLS models (Figure S3); ROC curves from SVM, RF, or PC-LDA models (Figure S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rpowers3@unl.edu. Tel: (402) 472-3039. Fax: (402) 472-9402.

ORCID

Robert Powers: 0000-0001-9948-6837

Author Contributions

T.V., P.S., and F.B. performed the experiments; T.V., Y.X., and R.P. designed the experiments; and T.V. and R.P. analyzed the data and wrote the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon the work supported by the National Science Foundation under Grant Number (1660921). This work was supported, in part, by funding from the Redox Biology Center (P30 GM103335, NIGMS) and the Nebraska Center for Integrated Biomolecular Communication (P20 GM113126, NIGMS). The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

■ REFERENCES

(1) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* **2006**, *78*, 4281−4290.

(2) Fiehn, O. Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155−171.

(3) Wishart, D. S. Metabolomics: applications to food science and nutrition research. *Trends Food Sci. Technol.* **2008**, *19*, 482−493.

(4) Ramirez, T.; Daneshian, M.; Kamp, H.; Bois, F. Y.; Clench, M. R.; Coen, M.; Donley, B.; Fischer, S. M.; Ekman, D. R.; Fabian, E.; Guillou, C.; Heuer, J.; Hogberg, H. T.; Jungnickel, H.; Keun, H. C.; Krennrich, G.; Krupp, E.; Luch, A.; Noor, F.; Peter, E.; Riefke, B.; Seymour, M.; Skinner, N.; Smirnova, L.; Verheij, E.; Wagner, S.; Hartung, T.; van Ravenzwaay, B.; Leist, M. Metabolomics in toxicology and preclinical research. *ALTEX* **2013**, *30*, 209−225.

(5) Kell, D. B.; Goodacre, R. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Discovery Today* **2014**, *19*, 171−182.

(6) Putri, S. P.; Nakayama, Y.; Matsuda, F.; Uchikata, T.; Kobayashi, S.; Matsubara, A.; Fukusaki, E. Current metabolomics: practical applications. *J. Biosci. Bioeng.* **2013**, *115*, 579−589.

(7) Goonewardena, S. N.; Prevette, L. E.; Desai, A. A. Metabolomics and atherosclerosis. *Curr. Atheroscler. Rep.* **2010**, *12*, 267−272.

(8) Vermeersch, K. A.; Styczynski, M. P. Applications of metabolomics in cancer research. *J. Carcinog.* **2013**, *12*, 9.

(9) Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discovery* **2016**, *15*, 473−484.

(10) Zhang, A.; Sun, H.; Wang, X. Emerging role and recent applications of metabolomics biomarkers in obesity disease research. *RSC Adv.* **2017**, *7*, 14966−14973.

(11) Xu, Y.; Correa, E.; Goodacre, R. Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection. *Anal. Bioanal. Chem.* **2013**, *405*, 5063−5074.

(12) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **2015**, *879*, 10−23.

(13) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608−D617.

(14) Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2007**, *2*, 2692−2703.

(15) Chan, E. C.; Koh, P. K.; Mal, M.; Cheah, P. Y.; Eu, K. W.; Backshall, A.; Cavill, R.; Nicholson, J. K.; Keun, H. C. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *J. Proteome Res.* **2009**, *8*, 352−361.

(16) Ellis, D. I.; Brewster, V. L.; Dunn, W. B.; Allwood, J. W.; Golovanov, A. P.; Goodacre, R. Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chem. Soc. Rev.* **2012**, *41*, 5706−5727.

(17) Horvath, C. *High-Performance Liquid Chromatography: Advances and Perspectives*; Academic Press, 1980, pp 91−108.

(18) Harris, D. *Quantitative Chemical Analysis*, 7th ed.; Craig Bleyer: Michelson Laboratory China Lake, California, 2007; p 828.

(19) Michael Borgerding, T. P. Determination of Nicotine in Tobacco, Tobacco Processing Environments and Tobacco Products. In *Analytical Determination of Nicotine and Related Compounds and their Metabolites*; Gorrod, J. W., Jacob, P., Eds.; Elsevier, 1999; pp 285−392.

(20) Amberg, A.; Riefke, B.; Schlotterbeck, G.; Ross, A.; Senn, H.; Dieterle, F.; Keck, M. NMR and MS Methods for Metabolomics. *Methods Mol. Biol.* **2017**, *1641*, 229−258.

(21) Worley, B.; Powers, R. Multivariate Analysis in Metabolomics<tep-common:author-query>AQ3: Please provide a DOI number for ref 21 or indicate if one doesn&amp;#x2019;t exist.</tep-common:author-query>. *Curr. Metabolomics* **2012**, *1*, 92−107.

(22) Saccenti, E.; Hoefsloot, H. C. J.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M. W. B. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014**, *10*, 361−374.

(23) Bhinderwala, F.; Wase, N.; DiRusso, C.; Powers, R. Combining Mass Spectrometry and NMR Improves Metabolite Detection and Annotation. *J. Proteome Res.* **2018**, *17*, 4017−4022.

(24) Marshall, D. D.; Powers, R. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, *100*, 1−16.

(25) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *London, Edinburgh Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559−572.

(26) Nicholson, J. K.; Lindon, J. C.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29*, 1181−1189.

(27) Worley, B.; Powers, R. PCA as a practical indicator of OPLS-DA model reliability. *Curr. Metabolomics* **2016**, *4*, 97−103.

(28) Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166−173.

(29) Trevor Hastie, R. T. Jerome Friedman. In *The Elements of Statistical Learning*, 2nd ed.; Springer, 2009; pp 106−117.

(30) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20*, 341−351.

(31) Yang, J.; Yang, J.-Y. Why can LDA be performed in PCA transformed space? *Pattern Recognit.* **2003**, *36*, 563−566.

(32) Ingo Steinwart, A. C. *Support Vector Machines*; Springer Science +Business Media, LLC, 2008; p 611.

(33) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(34) Worley, B.; Powers, R. MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chem. Biol.* **2014**, *9*, 1138−1144.

(35) R core development team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.

(36) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; Dame, Z. T.; Poelzer, J.; Huynh, J.; Yallou, F. S.; Psychogios, N.; Dong, E.; Bogumil, R.; Roehring, C.; Wishart, D. S. The human urine metabolome. *PLoS One* **2013**, *8*, No. e73076.

(37) Hollas, J. M. *Modern Spectroscopy*, 4nd ed.; John Wiley & Sons, Ltd, 2004; pp 35−36.

(38) Nguyen, B. D.; Meng, X.; Donovan, K. J.; Shaka, A. J. SOGGY: Solvent-optimized double gradient spectroscopy for water suppression. A comparison with some existing techniques. *J. Magn. Reson.* **2007**, *184*, 263−274.

(39) Dunn W, W. S. *Pattern Recognition Techniques in Drug Design*; Pergamon Press: Oxford, 1990; pp 691−714.

(40) Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175−85.

(41) Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* **1999**, *10*, 988−999.

(42) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. In *A Training Algorithm for Optimal Margin Classifiers*, COLT '92 Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992; pp 144−152.

(43) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer, 2016; p 297.

(44) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179−188.

(45) Provost, F.; Fawcett, T.; Kohavi, R. In *The Case Against Accuracy Estimation for Comparing Induction Algorithms*, The Fifteenth International Conference on Machine Learning, 1997.

(46) Lusted, L. B. Logical analysis in roentgen diagnosis. *Radiology* **1960**, *74*, 178−193.

(47) Kjeldahl, K.; Bro, R. Some common misunderstandings in chemometrics. *J. Chemom.* **2010**, *24*, 558−564.

(48) Eriksson, L.; Trygg, J.; Wold, S. CV-ANOVA for significance testing of PLS and OPLS models. *J. Chemom.* **2008**, *22*, 594−600.

(49) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA cross validation. *Metabolomics* **2008**, *4*, 81−89.

(50) Han, H.; Jiang, X. Overcome support vector machine diagnosis overfitting. *Cancer Inf.* **2014**, *13*, No. CIN-S13875.