Research Article

Received: 10 March 2017 Revised: 25 May 2017 Accepted article published: 5 June 2017 Published online in Wiley Online Library: 28 July 2017

(wileyonlinelibrary.com) DOI 10.1002/jsfa.8462

Comparison of partial least squares and random forests for evaluating relationship between phenolics and bioactivities of *Neptunia oleracea*

Soo Yee Lee,[a] Ahmed Mediani,[b] Maulidiani Maulidiani,[a] Alfi Khatib,[a,c] Intan Safinar Ismail,[a,d] Norhasnida Zawawi[b] and Faridah Abas[a,b*]

Abstract

BACKGROUND: *Neptunia oleracea* is a plant consumed as a vegetable and which has been used as a folk remedy for several diseases. Herein, two regression models (partial least squares, PLS; and random forest, RF) in a metabolomics approach were compared and applied to the evaluation of the relationship between phenolics and bioactivities of *N. oleracea*. In addition, the effects of different extraction conditions on the phenolic constituents were assessed by pattern recognition analysis.

RESULTS: Comparison of the PLS and RF showed that RF exhibited poorer generalization and hence poorer predictive performance. Both the regression coefficient of PLS and the variable importance of RF revealed that quercetin and kaempferol derivatives, caffeic acid and vitexin-2-*O*-rhamnoside were significant towards the tested bioactivities. Furthermore, principal component analysis (PCA) and partial least squares–discriminant analysis (PLS-DA) results showed that sonication and absolute ethanol are the preferable extraction method and ethanol ratio, respectively, to produce *N. oleracea* extracts with high phenolic levels and therefore high DPPH scavenging and α-glucosidase inhibitory activities.

CONCLUSION: Both PLS and RF are useful regression models in metabolomics studies. This work provides insight into the performance of different multivariate data analysis tools and the effects of different extraction conditions on the extraction of desired phenolics from plants.

Keywords: *Neptunia oleracea*; metabolomics; partial least squares; random forest; phenolics; extraction conditions

INTRODUCTION

High dietary intake of plant-based food is associated with reduced risk of several chronic diseases, such as cancer, diabetes and cardiovascular disease. The protective effects against these diseases are attributed to the biological activities offered by the phytochemical constituents present in the plant. The phytochemicals that can be found in various plants include phenolics, terpenoids, glucosinolates and alkaloids. These constituents have been reported to possess a wide range of bioactivities. For instance, phenolics are the major components contributing to the antioxidant capacity of spinach.[1] Glucosinolates and their hydrolysis products present in cruciferous vegetables exhibit antitumor activity.[2] Lupeol, a triterpene found in fruits and vegetables, possesses antiinflammatory and anticancer properties.[3] In view of the dependence on the phytochemical content for the biological activities of plant, there is interest in studies of the relationship between phytochemicals present and tested bioactivities. These studies would help identify the phytochemical markers contributing to the tested activities.

Metabolomics approaches have been widely used to determine the relationship between phytochemicals and biological activities. They combine the use of advanced analytical tools, such as nuclear magnetic resonance (NMR) and mass spectrometry (MS), with multivariate data analysis (MVDA). One of the regression models that is commonly used in these approaches is partial least squares (PLS). The PLS model has been successfully applied to identify the potential active metabolite corresponding to several biological activities, including antioxidant,[4–7] antiproliferative,[8] anti-inflammation,[9] α-glucosidase inhibition[10] and adenosine A1 receptor binding[11] activities. Recently, a machine learning method, random forests,[12] has drawn increased popularity as an

* *Correspondence to: F Abas, Laboratory of Natural Products, Institute of Bioscience, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia. E-mail: faridah_abas@upm.edu.my*

a *Laboratory of Natural Products, Institute of Bioscience, Universiti Putra Malaysia, Serdang, Selangor, Malaysia*

b *Department of Food Science, Faculty of Food Science and Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia*

c *Department of Pharmaceutical Chemistry, Faculty of Pharmacy, International Islamic University Malaysia, Kuantan, Pahang, Malaysia*

d *Department of Chemistry, Faculty of Science, Universiti Putra Malaysia, UPM, Serdang, Selangor, Malaysia*

application of MVDA tools in omics research. It has been applied mostly as a classification model for plant[13] and biofluid[14–16] metabolomics. For regression purposes, it has been used in a quantitative structure–activity relationship (QSAR) model for drug design[17,18] and for the correlation of transcriptomics and metabolomics data with potato quality traits.[19] However, its application as a regression model for investigating the relationship between bioactivity and phytochemical constituent in metabolomics study has not yet been reported.

Neptunia oleracea is a tropical plant that is consumed as a vegetable, especially in Southeast Asia. It is also used to relieve high fevers and to remove toxic elements from the body.[20] The extract of this plant possesses potential antioxidant and α-glucosidase inhibitory properties[21] and phenolic compounds have been suggested to be potential contributors to these bioactivities.[22] Although the recovery of bioactive metabolites from plants depends on the extraction conditions,[23,24] the appropriate extraction conditions for extracting the valuable phenolics from *N. oleracea* have not yet been explored. Hence, in order to obtain a high level of phenolics from this plant, the appropriate extraction conditions, such as solvent ratio and extraction method, should be determined.

The objectives of this study were to develop and compare the performance of PLS and RF regression models based on their goodness of prediction for the antioxidant and α-glucosidase inhibition of *N. oleracea* and to investigate the relationship between the phenolic constituents and tested bioactivities, hence highlighting the metabolites in the phenolic group that contributes to the bioactivities. Furthermore, different extraction conditions (solvent ratio and extraction method) were evaluated to identify suitable conditions that can yield a high level of those metabolites and thereby high bioactivities. The information obtained from this study may provide insight into the performance of different MVDA tools and facilitate optimization of the extraction method and ethanol ratio for preserving the desired phenolic compounds from plants.

EXPERIMENTAL

Chemicals

Gallic acid, quercetin, Folin–Ciocalteu reagent, sodium carbonate, phosphate buffer, α-glucosidase enzyme, *p*-nitrophenyl-α-D-glucopyranose (PNPG), glycine and 2,2-diphenyl-1-picrylhydrazyl (DPPH) were purchased from Sigma-Aldrich (Hamburg, Germany). Absolute ethanol, deuterated methanol-d_4 (CH_3OH-d_4), non-deuterated potassium phosphate monobasic (KH_2PO_4), sodium deuterium oxide (NaOD), trimethylsilyl propionic (TSP) acid-d4 sodium salt and deuterium oxide (D_2O) were supplied by Merck (Darmstadt, Germany).

Plant material

Neptunia oleracea was planted in Universiti Putra Malaysia Agricultural Park by distributing the stems in a pond. The plant was identified by an in-house botanist (Dr Shamsul Khamis) of the Institute of Bioscience and a voucher specimen (SK2516/14) was provided.

Sampling and sample preparation

For harvesting and sampling, the plot was divided into six individual subplots for biological replications. During the harvesting period, the leaves were separated from the stems. Immediately, the leaves were cleaned and stored in a deep freezer at −80 °C overnight, followed by freeze-drying. The drying process was complete when the weight of the leaf remained constant. Dried samples were ground into a fine powder using a laboratory blender. The samples were stored in aluminium pouches to avoid exposure to light and atmospheric moisture.

Extraction

Powdered leaf samples were subjected to two different extractions methods (sonication and soaking). For each extraction method, three different ethanol ratios (50%, 80% and absolute) were used. Thus 36 crude extracts were prepared. Soaking was carried out by immersing 2 g ground samples in 100 mL solvent in a conical flask at room temperature for 5 days. The same procedure was followed for sonication, except that the mixture was subjected to sonication (at a controlled temperature) in an ultrasonic bath sonicator (Branson, 141 8510E-MTH model, Danbury, USA) for 1 h. All the mixtures were transferred to Nalgene polypropylene copolymer centrifuge bottles (NY, USA) and centrifuged at 13 000 rpm for 30 min to separate the supernatant and precipitates. The collected supernatant was then concentrated using a rotary evaporator and freeze dried to yield the crude extract. The crude extracts were stored at 4 °C until further analysis.

Total phenolic content (TPC) determination

TPC was determined using Folin–Ciocalteu reagent as previously reported.[21] Samples were prepared at 0.5 mg mL^{-1} by dissolving the extracts in DMSO. A total of 20 µL of the samples was mixed with 100 µL Folin–Ciocalteu reagent in 96-well plates. After 5 min of incubation, 80 µL of 7.5% sodium carbonate solution was added. The microplate was then covered and incubated in the dark for 30 min. The absorbance was measured at 765 nm using a SPECTRAmax PLUS microplate reader (Molecular Devices, CA, USA). Each sample was analyzed in three replicates. A standard curve of gallic acid was generated for the total phenolic content calculations, and the results were expressed in micrograms of gallic acid equivalents per milligram of extract (µg GAE mg^{-1} extract).

DPPH free radical scavenging assay

The DPPH free radical scavenging assay was performed as previously described.[21] Samples were prepared by dissolving 1 mg of the extracts in 1 mL DMSO, followed by dilution with the same solvent to reach final concentrations ranging from 50 to 0.4 µg mL^{-1}. To 50 µL of test samples loaded into each well, 100 µL DPPH (5.9 mg 100 mL^{-1} methanol) was added and mixed well. The mixture was then incubated in the dark for 30 min. The absorbance of the mixtures was measured at 517 nm using a SPECTRAmax PLUS microplate reader. The scavenging capacity (SC) was calculated as $\%SC = [(A_o - A_s)/A_o] \times 100\%$, where A_o and A_s are the absorbance values of the reagent blank and tested samples, respectively. Each sample was analyzed in three replicates. The results were expressed as the IC_{50} values (µg mL^{-1}), which corresponded to the sample concentration required to scavenge 50% of the DPPH free radicals. Quercetin was used as positive control in the assay.

α-Glucosidase inhibition assay

The α-glucosidase inhibition assay was conducted according to an established method, with minor modifications.[21] The α-glucosidase enzyme and PNPG substrate were separately prepared in 50 mmol L^{-1} phosphate buffer (pH 6.5). The α-glucosidase enzyme was diluted to obtain a final concentration of 0.02 U in

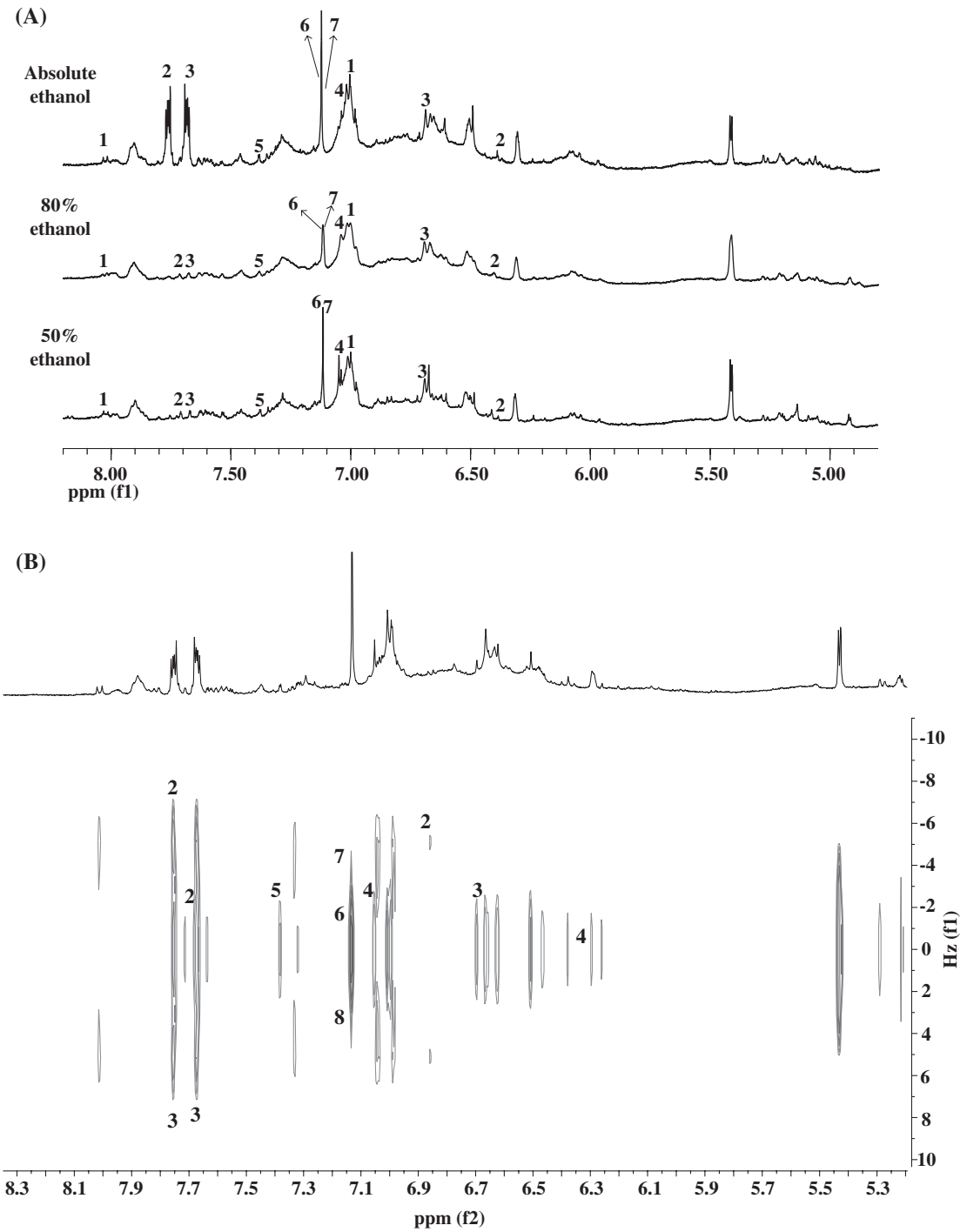

Figure 1. Representative ^{1}H NMR spectra of three different ethanol ratios (A) and two-dimensional *J*-resolved spectra (B) of *N. oleracea* leaf extracts from 5.0 to 8.0 ppm. Signal assignment for phenolic constituents: 1, vitexin-2-*O*-rhamnoside; 2, quercetin derivatives; 3, kaempferol derivatives; 4, myricetin derivatives; 5, catechin; 6, caffeic acid; 7, gallic acid; 8, 3,4-*O*-dimethylgallic acid.

each well, and the PNPG substrate was prepared at a concentration of 1 mmol L^{-1}. Samples were prepared in DMSO at a concentration of 1 mg mL^{-1} as stock solutions and diluted with 30 mmol L^{-1} phosphate buffer to reach a final concentration of 5 µg mL^{-1}. The samples were then diluted twofold using DMSO containing buffer to obtain a series of concentrations ranging from 5 to 0.04 µg mL^{-1}. The final concentration of DMSO in each well was 0.5%. To test the enzyme inhibition by the sample extracts, each well was loaded with 130 µL of 30 mmol L^{-1} phosphate buffer, followed by 10 µL sample and 10 µL of the enzyme. The extract was allowed to interact with the enzyme for 5 min at room temperature before the reaction was initiated by adding 50 µL of the substrate. The total volume in each well was 200 µL. After 15 min incubation at room temperature, the reaction was stopped by adding 50 µL of 2 mol L^{-1} glycine (pH 10). The microplate was read with a SPECTRAmax PLUS spectrophotometer at 405 nm. The percentage of inhibition was calculated as percent inhibition = $[(a_n - a_s)/a_n] \times 100\%$, where a_n and a_s are the absorbance values of the negative control and

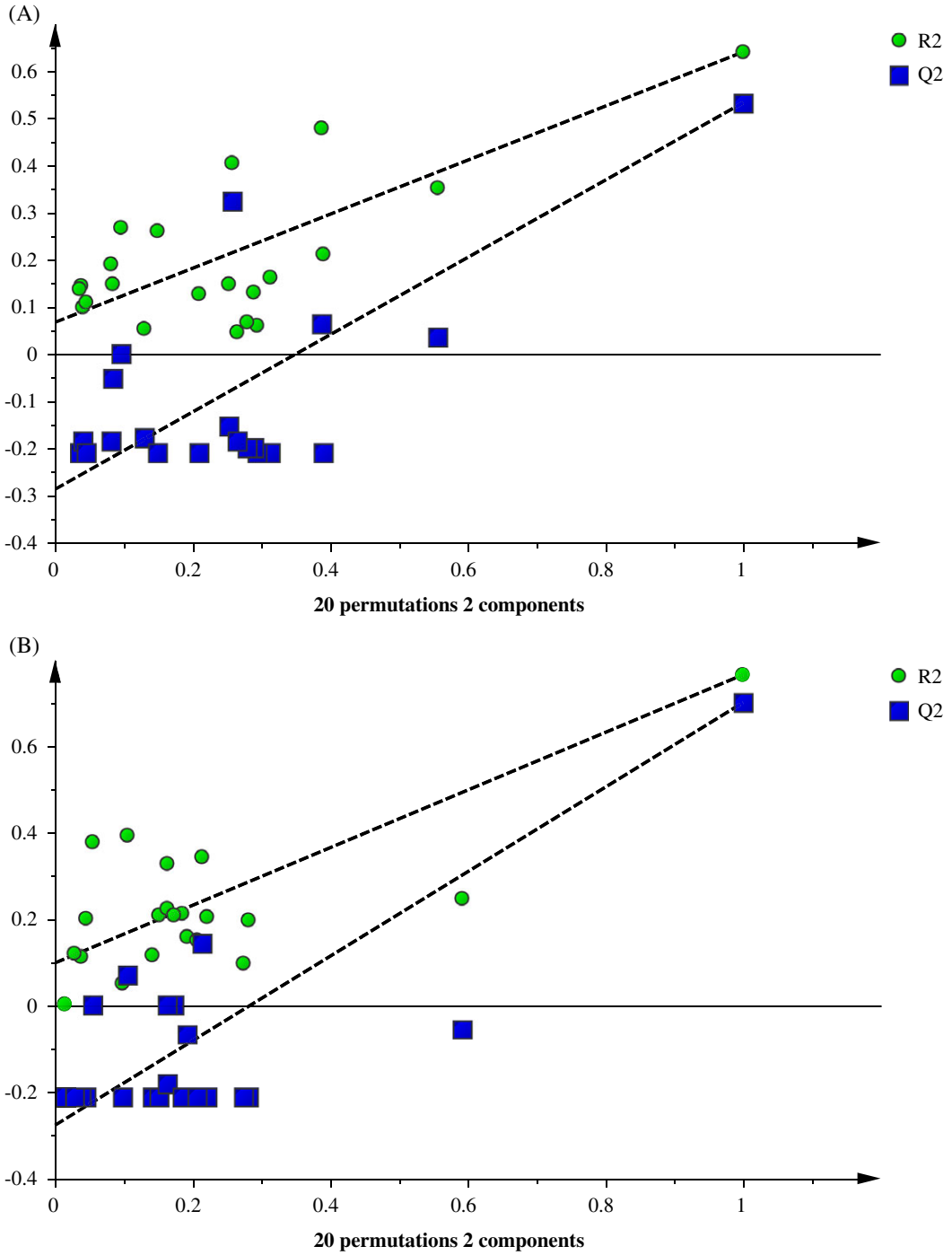

Figure 2. Permutation plots of PLS model describing the R^2 and Q^2 *Y*-intercepts for DPPH free radical scavenging (A) and α-glucosidase inhibitory (B) activities of *N. oleracea*.

tested samples, respectively. Each sample was analyzed in three replicates. The results were expressed as IC_{50} values (μg mL^{-1}). Quercetin was used as positive control in the assay.

NMR measurements

Sample preparation and NMR measurements were carried out as previously described.[25] Ten milligrams of crude extract was mixed with 0.375 mL CH_3OH-d_4 and 0.375 mL KH_2PO_4 in D_2O (pH 6.0, containing 0.1% TSP). The mixtures were ultrasonicated for 15 min and then centrifuged at 13 000 rpm for 10 min to separate the supernatant from the residue. Subsequently, 0.6 mL of the supernatant was transferred to an NMR tube for ^{1}H NMR analysis at 26 °C. A 500 MHz Varian INOVA NMR spectrometer operating at a frequency of 499.887 MHz was used for the analysis. The acquisition time for each ^{1}H NMR spectrum was 3.54 min, and 64 scans were performed. Chenomx software v. 5.1 (Edmonton, Canada) was used to correct the phasing and baselines of all the NMR spectra. TSP was used as an internal standard, and the spectra were manually normalized to TSP. Two-dimensional *J*-resolved NMR spectra were collected to provide additional support for the assignment and confirmation of some compounds.

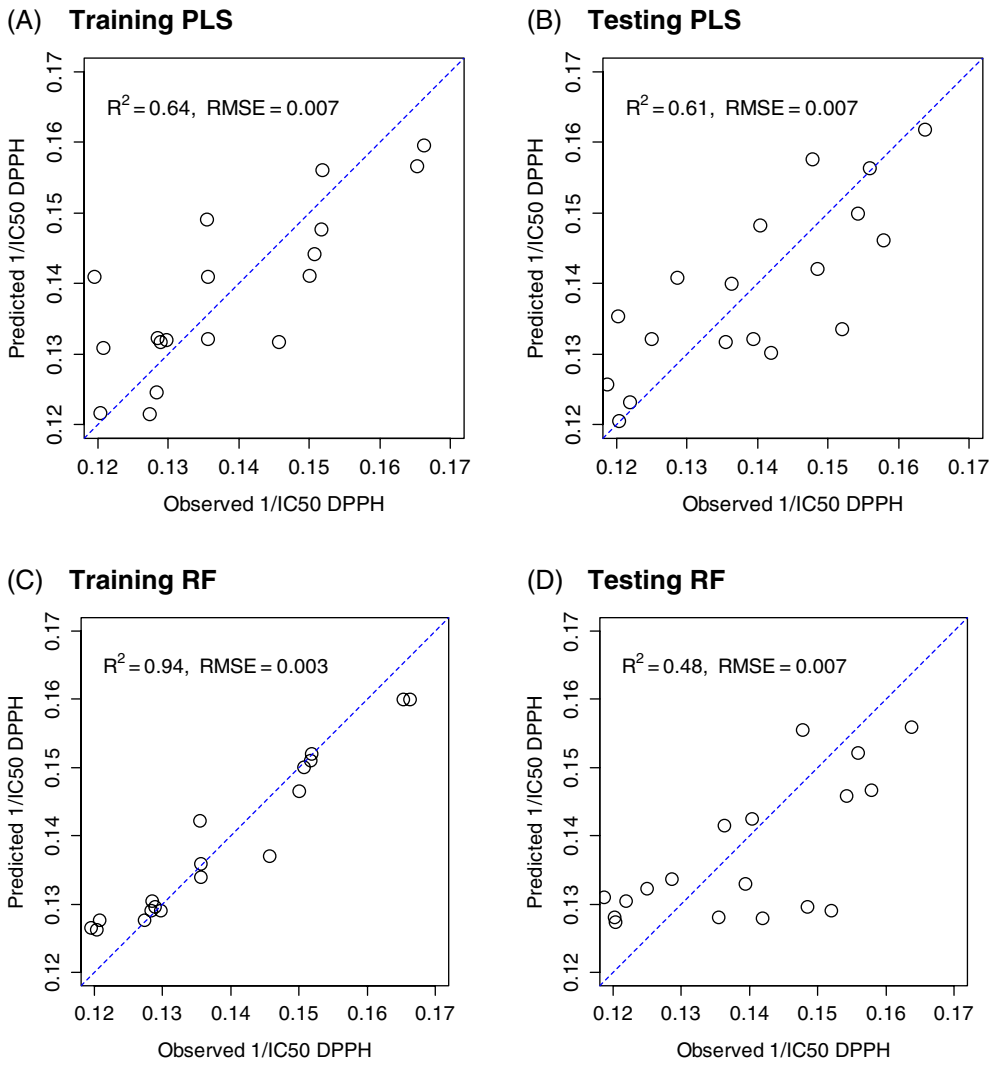

Figure 3. Scatter plots describing the relationship between observed and predicted $1/IC_{50}$ of DPPH free radical scavenging activity for training PLS (A), testing PLS (B), training RF (C) and testing RF (D). Dashed lines represent the concordance lines $y = x$.

Identification of phenolic constituents

In our previous study,[22] several phenolic compounds were found to be potential contributors to the antioxidant and α-glucosidase inhibitory activities of *N. oleracea*. These phenolic constituents included caffeic, gallic and 3,4-*O*-dimethylgallic acids, vitexin-2-*O*-rhamnoside, catechin and derivatives of quercetin, kaempferol and myricetin. The NMR signals of these phenolics were also detected in this study. The representative ^{1}H NMR spectra of *N. oleracea* extracts from three different ethanol ratios and the assignment of the phenolics are shown in Fig. 1. The signals of these identified phenolics were used in the PLS and RF regression models and other analyses in this study. The information regarding the characteristic NMR signals of each of the identified phenolics can refer to the previous report.[22]

Bucketing of ^{1}H NMR spectra

The ^{1}H NMR spectra were processed, bucketed and converted to ASCII files using Chenomx software version 6.2. A total of 245 integrated regions were obtained by binning the δ 0.5–10.0 region with a width of δ 0.04. The water and methanol signals at δ 4.70–4.88 and δ 3.27–3.35, respectively, were excluded.

Development of regression models

PLS model

The PLS model was developed using SIMCA-P software version 13.0 (Umeå, Sweden). The ^{1}H NMR chemical shifts of the identified phenolics were the *X* variables, whereas the DPPH scavenging and α-glucosidase inhibitory activities ($1/IC_{50}$ values) were the *Y* variables. A total of 36 observation datasets were input into the software. The scaling method applied was Pareto scaling. The datasets were randomly separated into two groups, with 50% for both training and testing. The training dataset was used to develop the model, whereas the testing dataset was used to validate the model.

RF model

The RF model was developed using the Random Forest Package[26] in R software. Unlike the standard classification regression tree (CRT), which is established by a single decision tree, RF grows multiple trees, similar to forests. Some observation and variable data become subsets randomly during trees developing using the bootstrapping technique. In the case of regression, the average of the output value from the trees is taken for the prediction. According to a compromise between the accuracy and processing time, Oshiro *et al.*[27] recommended the number of trees to be

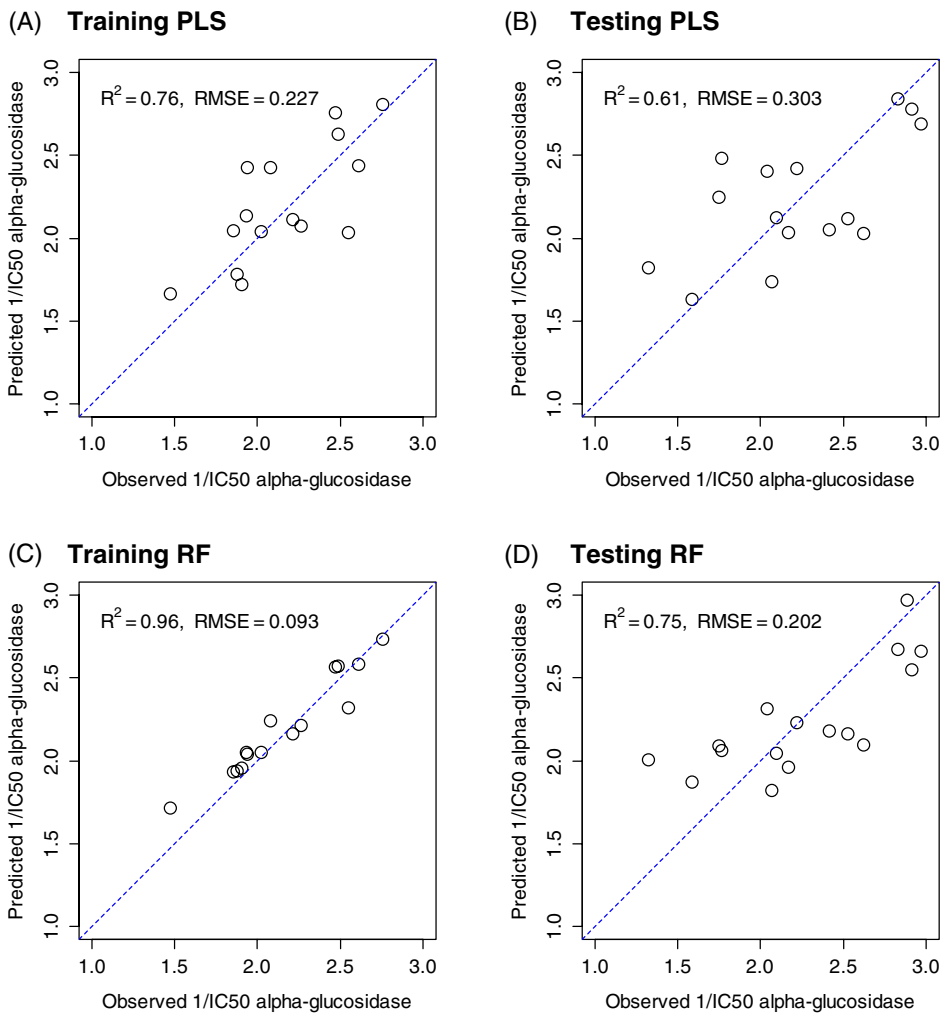

Figure 4. Scatter plots describing the relationship between observed and predicted $1/IC_{50}$ of α-glucosidase inhibitory activity for training PLS (A), testing PLS (B), training RF (C) and testing RF (D). Dashed lines represent the concordance lines $y = x$.

between 64 and 128. Hence 100 trees were used in this study. The *X* and *Y* variables were the same as in the PLS model.

Pattern recognition analysis

Pattern recognition analysis was used to evaluate the phenolic variation among *N. oleracea* leaf extracts obtained using different extraction methods and ethanol ratios, and hence the effect of the different extraction conditions. The pattern recognition methods used in this study were principal component analysis (PCA) and partial least squares–discriminant analysis (PLS-DA). Both analyses were performed using SIMCA-P software version 13.0. The ^{1}H NMR chemical shifts of the identified phenolics were the input X variables. Pareto scaling method was applied.

Statistical analysis

Minitab software version 17 (Minitab Inc., State College, PA, USA) and InStat version 2.02 statistical package (GraphPad Software, San Diego, CA, USA) were used to analyze the TPC, DPPH scavenging and α-glucosidase inhibition assay data. The results were expressed as the mean ± SD of six biological replicates. To determine the significant differences, analysis of variance (ANOVA) was applied. Values were considered to differ significantly when the *P*-value was less than 0.05.

RESULTS AND DISCUSSION

Performance of regression models

Model validation

Validation of the developed PLS and RF regression models was performed before they were compared. The PLS model was validated using internal cross-validation by means of cumulative R^2 and Q^2, permutation test and external validation. Cumulative R^2 and Q^2 indicate goodness of fit and predictive ability of the model, respectively. The criteria to be a good model include $Q^2 > 0.5$, $R^2 > Q^2$ and the difference between these values being within the range 0.2–0.3. In this study, autofit of the PLS model in SIMCA resulted in two components, with R^2 and Q^2 values of 0.807 and 0.656, respectively. This showed that the PLS model meets the criteria for validation and prediction performance. The permutation test is another commonly used validation approach in metabolomics study. It provides an unbiased assessment of the validity and degree of overfitting of the PLS model by comparing the R^2 and Q^2 of the original model with those of the models where the *Y* variable has been permuted randomly. The measure of overfitting is indicated by the intercepts of R^2 and Q^2.[28] Figure 2 shows the permutation test of the current PLS model. The *Y*-intercepts of R^2 and Q^2 were less than 0.3 and 0.05, respectively, further showing that the model was safe from overfitting.[28] External

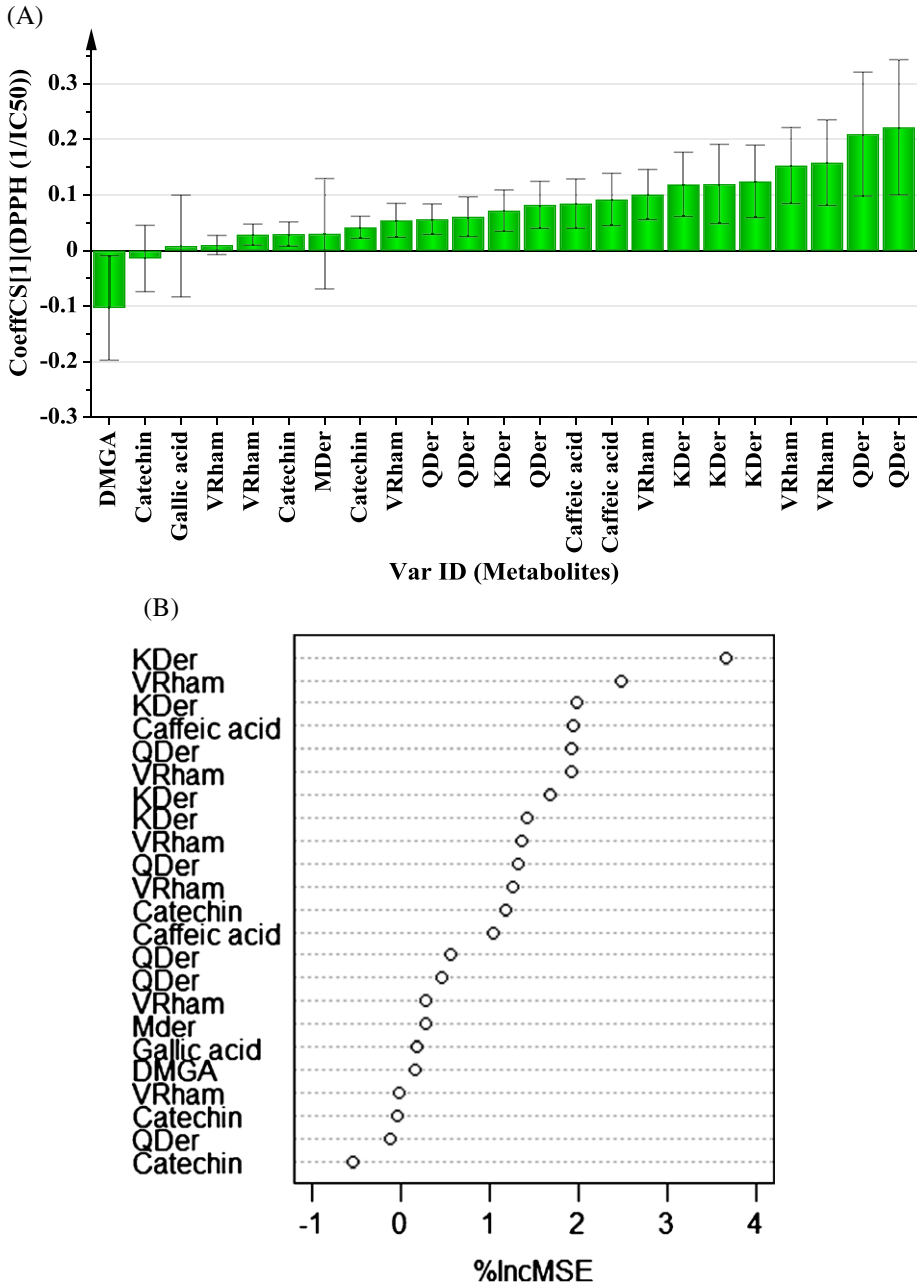

Figure 5. Coefficient plot of PLS (A) and variable importance of random forest (B) in contribution to DPPH free radical scavenging activity. Key: DMGA, 3,4-*O*-dimethylgallic acid; VRham, vitexin-2-*O*-rhamnoside; MDer, myricetin derivatives; QDer, quercetin derivatives; KDer, kaempferol derivatives.

validation using the testing datasets also showed that the PLS model was good based on the R^2 at 0.61 for both DPPH free radical scavenging and α-glucosidase inhibitory activities in the testing sets (Figs 3B and 4B).

Since the RF takes the randomness out of data in observation and variables, it yielded not exactly the same (but still similar) prediction of the DPPH scavenging and α-glucosidase inhibitory activities ($1/IC_{50}$ values) when it was rerun. Thus RF was run 100 times, and the average was taken to predict the two bioactivities. The RF model was validated using the testing datasets. Validation of the model yielded a relatively low R^2 value of 0.48 for DPPH free radical scavenging (Fig. 3D) but resulted in high agreement for the observed α-glucosidase inhibitory activity ($R^2 = 0.75$) (Fig. 4D). The smaller value of R^2 for DPPH compared to that for α-glucosidase inhibitory may be due to the narrow range of values of $1/IC_{50}$ of DPPH (0.11–0.17).

Comparison of PLS and RF models

The PLS and RF models were compared based on their performance to predict the DPPH free radical scavenging and α-glucosidase inhibitory activities of *N. oleracea* extracts. The predictive performance of the model can be revealed by the relationship between observed and predicted values of the bioactivities. Good agreement between the observed and predicted bioactivities suggests that the model is reliable in predicting the bioactivities of new samples based on their ^{1}H NMR data. Conformity between the observed and predicted bioactivities can be evaluated based on the root mean square error (RMSE) and R^2

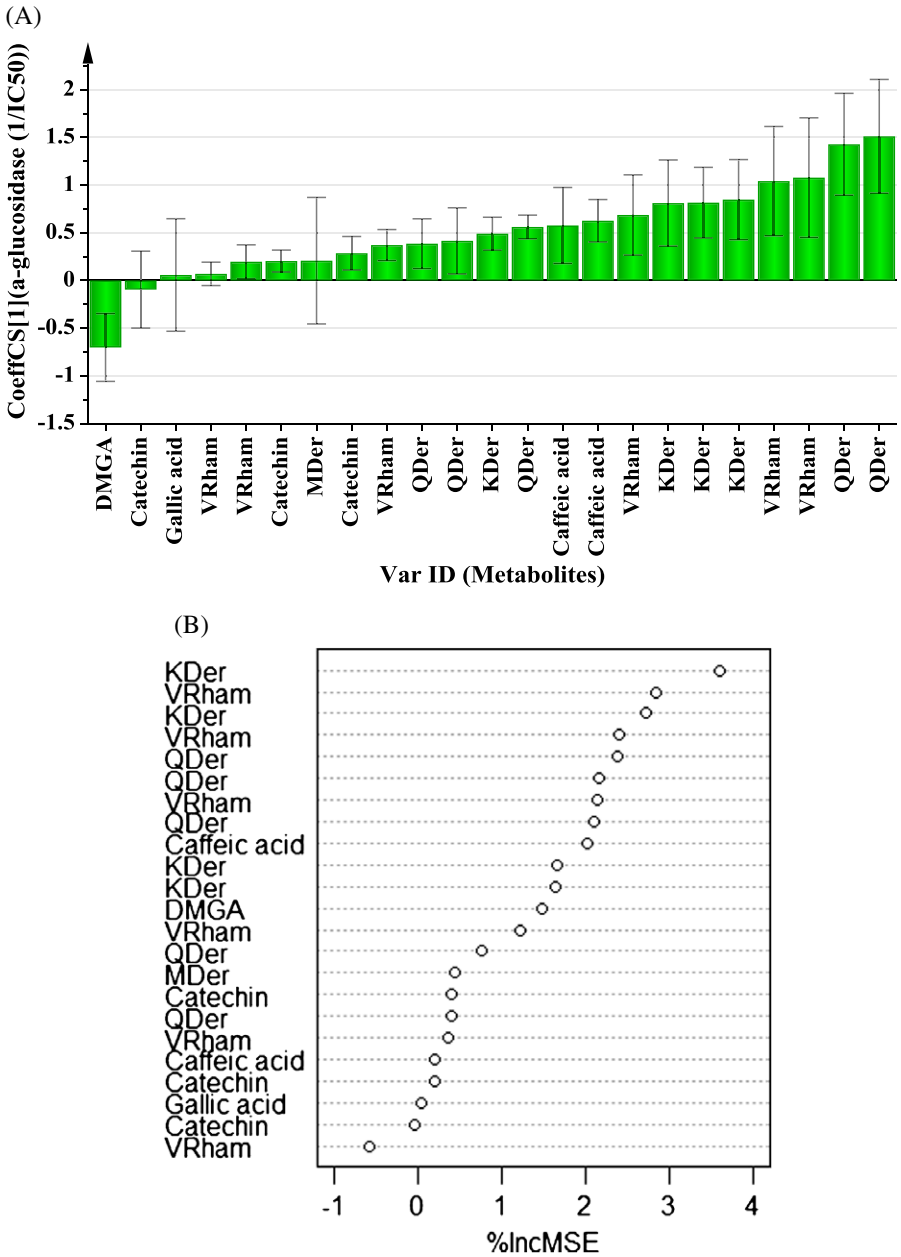

Figure 6. Coefficient plots of PLS (A) and variable importance of random forest (B) in contribution to α-glucosidase inhibitory activity. Key: DMGA, 3,4-*O*-dimethylgallic acid; VRham, vitexin-2-*O*-rhamnoside; MDer, myricetin derivatives; QDer, quercetin derivatives; KDer, kaempferol derivatives.

values. A lower value of RMSE indicates good conformity, while R^2 is the opposite of RMSE.

Figure 3 presents scatter plots describing the relationship between observed and predicted DPPH free radical scavenging activity of the PLS and RF models. Figure 3(A, B) shows the results in the PLS training and testing datasets, respectively. The low RMSE and high R^2 values in all datasets of PLS revealed good conformity between the observed and predicted DPPH free radical scavenging activity. In addition, the similar RMSE and R^2 for the training ($R^2 = 0.64$ and RMSE = 0.007) and testing ($R^2 = 0.61$ and RMSE = 0.007) datasets in the PLS model showed that this model exhibited a good generalization and was safe from overfitting, and hence will generate reliable bioactivity results for the new samples based on their ^{1}H NMR data. On the other hand, Fig. 3(C, D) displays the results in the RF training and testing datasets, respectively. The training dataset of the RF model showed excellent conformity between the observed and predicted DPPH free radical scavenging activity with the large R^2 value ($R^2 = 0.94$). However, the R^2 value decreased drastically in the testing dataset ($R^2 = 0.48$). The inconsistent results in the training and testing datasets showed that the RF fitted well only in the training dataset but fitted poorly in the testing dataset. This indicated the weak generalization and high tendency of overfitting, and therefore poor predictive performance of the RF model. The results for the observed and predicted α-glucosidase inhibitory activity showed the same trend as that for the DPPH free radical scavenging activity (Fig. 4).

These findings showed that the PLS model resulted in better predictive performance in contrast to the RF model. The excellent predictive performance of PLS was in good agreement with the

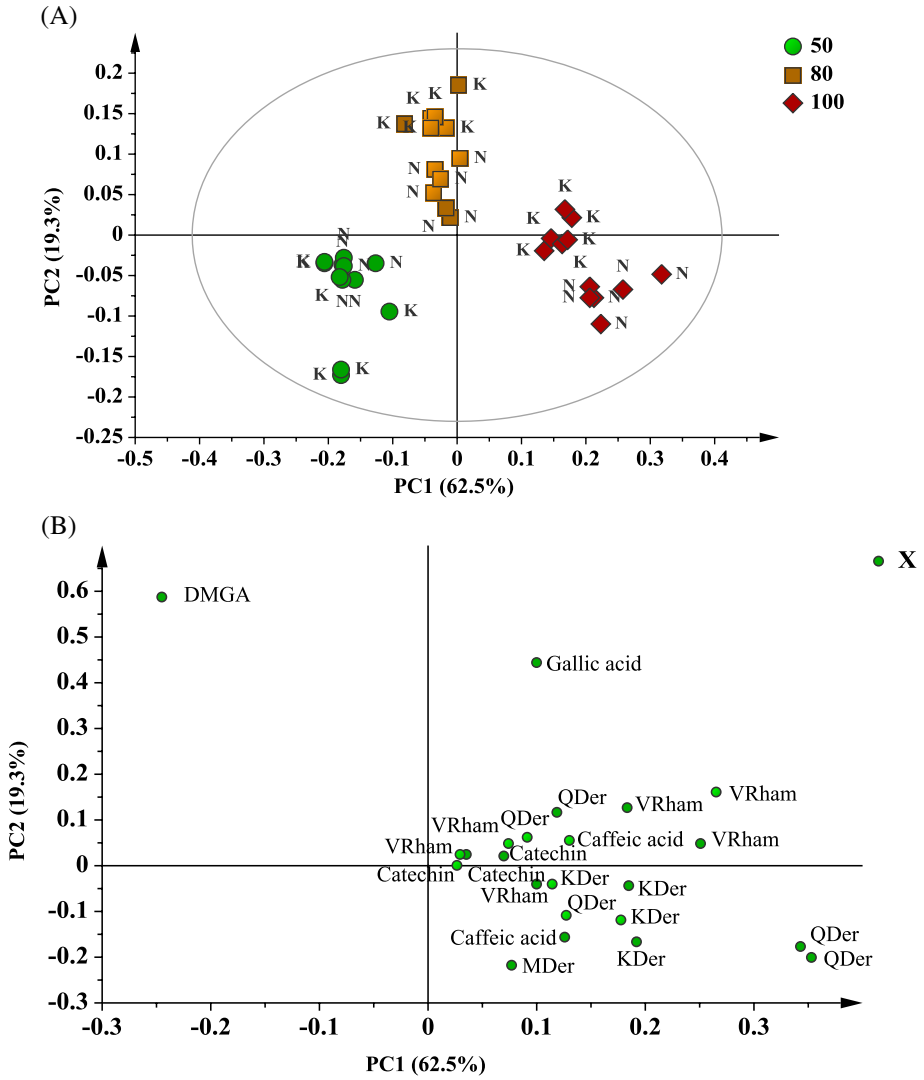

Figure 7. Score and loading plots of PCA for evaluating effect of different extraction conditions (extraction method and ethanol ratio) on phenolic constituent levels. Key: N, sonication; K, soaking; DMGA, 3,4-*O*-dimethylgallic acid; VRham, vitexin-2-*O*-rhamnoside; MDer, myricetin derivatives; QDer, quercetin derivatives; KDer, kaempferol derivatives.

previous report regarding the outperformance of PLS as compared to a nonlinear model: artificial neural network (ANN) in predicting the antioxidant activity of *Pegaga* extracts.[28] As for RF, although the model showed a higher tendency of overfitting, its validity was still relatively high in the agreement between the observed and predicted α-glucosidase inhibitory activity (Fig. 4B, C). Thus RF might be suitable for the prediction of α-glucosidase inhibitory activity of *N. oleracea*, but not for the prediction of DPPH free radical scavenging activity. For classification, Gromski *et al.*[29] explained that RF is robust against overfitting. However, according to the findings presented in this work, RF showed a high degree of overfitting in the case of regression. Therefore, extra care should be taken to use RF as a regression model, and external cross-validation should be performed to check its degree of overfitting.

Relationship between phenolic constituents and bioactivities

Despite the relatively poorer predictive performance of RF, it was applied along with PLS to evaluate the relationship between the identified phenolics and the DPPH free radical scavenging and α-glucosidase inhibitory activities of *N. oleracea*. Results obtained from the two regression tools with different algorithms can provide more information regarding the contribution of the phenolics towards the studied bioactivities. This can help to identify the metabolites in the phenolic group that contribute to the bioactivities. The relationship between the identified phenolics and studied bioactivities were evaluated based on the regression coefficient of PLS and the variable importance of RF.

The regression coefficient of the metabolite signals shows how the phenolics affect the bioactivities. A high regression coefficient (positive or negative) indicates a greater effect on the bioactivities exhibited by the metabolite and vice versa. The regression coefficient plot of PLS for DPPH free radical scavenging is presented in Fig. 5(A). Almost all signals exhibited a positive regression coefficient for the DPPH free radical scavenging activity. This shows that they contributed to the DPPH free radical scavenging activity of *N. oleracea*. This is not surprising as phenolics have been highlighted as potential antioxidants in this plant.[22] Among the identified phenolics, quercetin and kaempferol derivatives, vitexin-2-*O*-rhamnoside and caffeic acid contribute the most towards the DPPH free radical scavenging activity of *N. oleracea*.

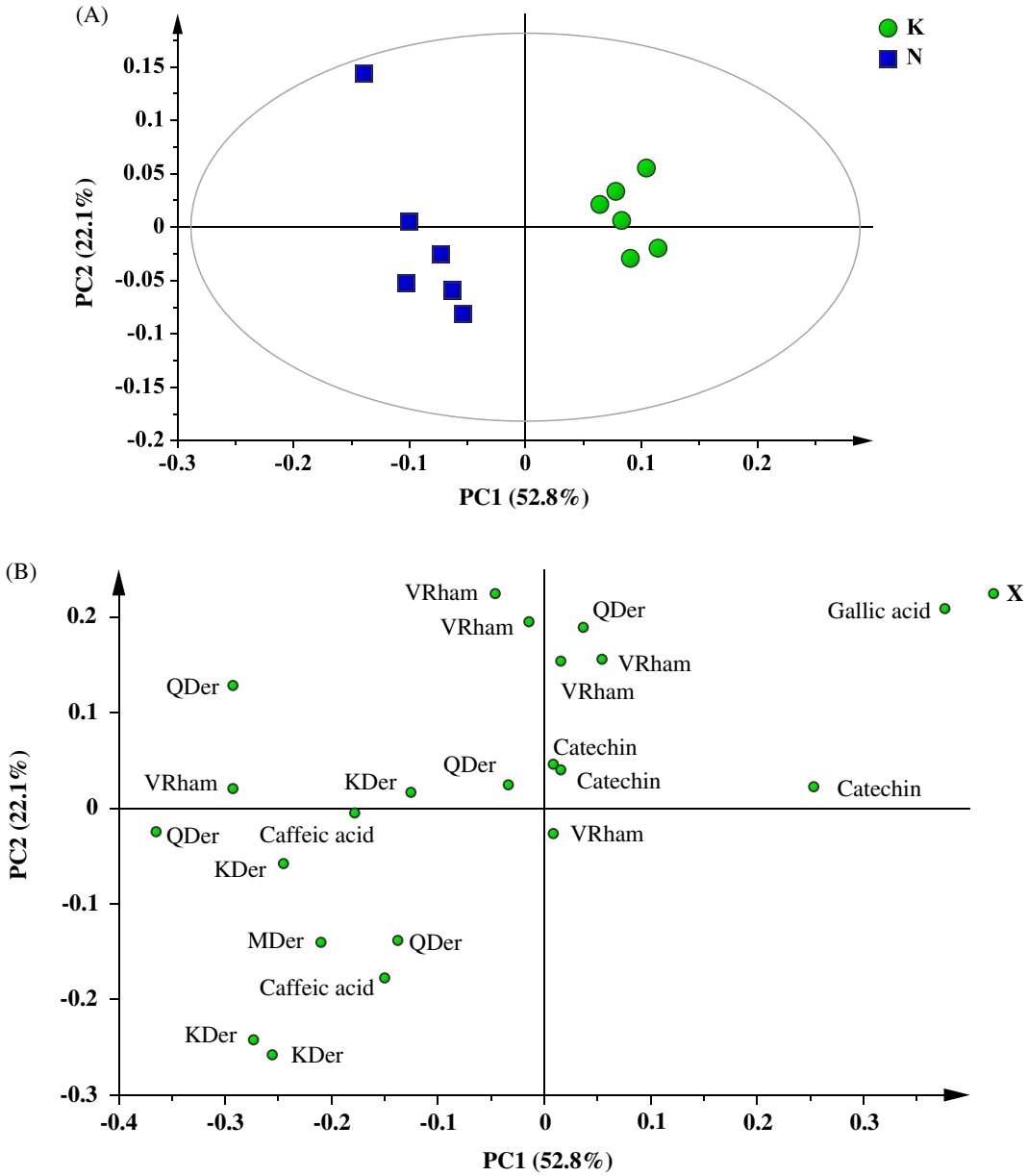

Figure 8. Score and loading plots of PLS-DA for evaluating phenolic variations among absolute ethanol abstracts obtained via two different extraction methods. Key: N, sonication; K, soaking; DMGA, 3,4-*O*-dimethylgallic acid; VRham, vitexin-2-*O*-rhamnoside; MDer, myricetin derivatives; QDer, quercetin derivatives; KDer, kaempferol derivatives.

The results for the α-glucosidase inhibitory activity were similar to those for the DPPH free radical scavenging activity (Fig. 6A).

The variable importance of RF is revealed by the percent increase in the mean square error (%IncMSE). It is obtained from the average value of %IncMSE of variable importance from 100 times running of RF. The higher the value, the more important is the variable and vice versa. Figure 5(B) shows the variable importance of the phenolics to the DPPH free radical scavenging activity of *N. oleracea* based on the RF model. Similar to the regression coefficient of PLS, the results of variable importance of RF revealed that quercetin and kaempferol derivatives, vitexin-2-*O*-rhamnoside, and caffeic acid were the most important metabolites among the identified phenolics regarding the contribution to the DPPH free radical scavenging activity of *N. oleracea*. Similar results were also observed for the α-glucosidase inhibitory activity (Fig. 6B).

The consistency of the results from PLS and RF strongly suggests that quercetin and kaempferol derivatives, vitexin-2-*O*-rhamnoside, and caffeic acid are important phytochemical markers of the DPPH scavenging and α-glucosidase inhibitory activities of *N. oleracea*. The similar response by these phenolics towards the two studied bioactivities also showed that they are both strong DPPH free radical scavengers and potent α-glucosidase inhibitors. These results agree with previous results that reported the potent antioxidant and α-glucosidase inhibitory activities of quercetin and kaempferol, as well as their derivatives.[30–32] Moreover, based on our previous findings,[22] the derivatives of quercetin and kaempferol that are present in *N. oleracea* were quercetin-3-*O*-arabinoside, quercetin-3-*O*-rhamnoside, quercetin, rutin, kaempferol-3-*O*-glucoside and kaempferol-3-*O*-rhamnoside. Furthermore, caffeic acid and vitexin-2-*O*-rhamnoside have also

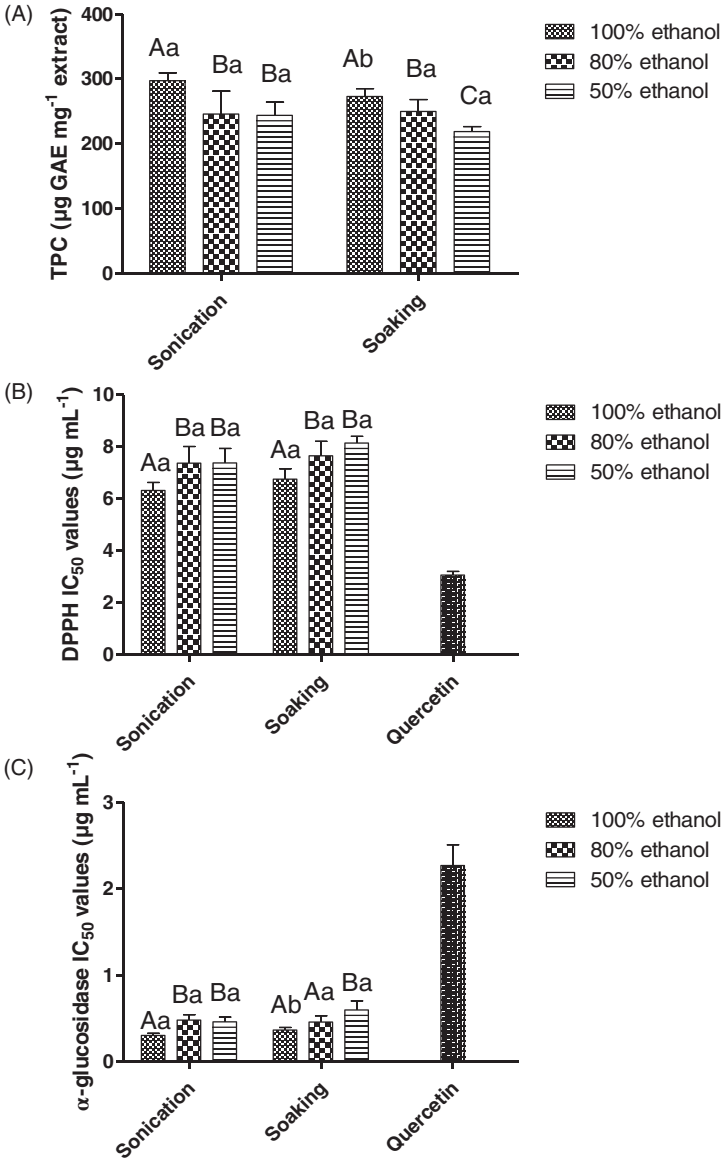

Figure 9. Statistical analysis on total phenolic content (A), DPPH IC_{50} (B) and α-glucosidase IC_{50} (C) of *N. oleracea* leaf extracts by different extraction methods and ethanol ratios. The first, upper-case letter refers to the comparison of different solvent ratios for the same extraction method. The second, lower-case letter refers to the comparison between different extraction methods for the same solvent ratio. Means with different letters are significantly different ($P < 0.05$).

been reported to be potential antioxidant and α-glucosidase inhibitors.[23,33–35] The results of this present study reveal that the phenolics in *N. oleracea* are potent antioxidants and α-glucosidase inhibitors. In addition, this study also demonstrated that PLS and RF are not only useful in predicting the bioactivities but also helpful in highlighting the important phytochemical markers.

Effect of different extraction conditions (extraction method and ethanol ratio) on phenolics

The effect of different extraction methods and ethanol ratios on the phenolic constituent levels was evaluated in order to identify the extraction conditions for producing *N. oleracea* extract with high levels of quercetin and kaempferol derivatives, vitexin-2-*O*-rhamnoside and caffeic acid. PCA was used for this purpose. PCA is an unsupervised MVDA that is used to classify samples according to their chemical composition.[36] The score plot of PCA provides an understanding of the clustering features of the samples, whereas the loading plot highlights the compounds responsible for the discrimination. As shown by the score plot in Fig. 7(A), the *N. oleracea* extracts obtained from the different extraction methods and ethanol ratios were separated into three clusters with no notable outliers, where PC1 and PC2 contributed to 62.5% and 19.3% of the variance, respectively. It can be seen that the separation of the samples was more attributable to the ethanol ratios than to the extraction methods. The extracts produced using three different ethanol ratios were clearly separated into three clusters, but the samples obtained via the two different extraction methods were not separated from one another, particularly in the 50% ethanol extracts. However, the separation between the two extraction methods increased as the ethanol ratio increased. It can be observed that the two

extraction methods were slightly separated in the 80% ethanol extracts, and the separation became more obvious in the absolute ethanol extracts. Hence these results showed that the different extraction methods did not give substantial variation in the phenolic content, but varying the ethanol ratio did. The extraction method only gave variation when a high ethanol ratio was used.

The PCA loading plot (Fig. 7B) reveals that all of the identified phenolics, except for 3,4-*O*-dimethylgallic acid, were more abundant in the absolute ethanol extracts than in the 50% and 80% ethanol extracts. However, these phenolics were not discriminating the absolute ethanol extracts produced by the two different extraction methods. In order to maximize the separation between the absolute ethanol extracts obtained via sonication and soaking, and to reveal the underlying phenolic variation, PLS-DA, which is a supervised MVDA, was applied. PLS-DA separates the samples by rotating the PCA components to achieve the maximum separation. The resulting PLS-DA model was validated by cross-validation through cumulative R^2 and Q^2. The model showed high discrimination, with R^2 and Q^2 values of 0.97 and 0.95, respectively. The small *P*-value ($P = 0.003$) obtained via CV-ANOVA also revealed that the model was statistically significant. The PLS-DA score plot (Fig. 8A) showed that the absolute ethanol extracts obtained from the two different extraction methods were clearly separated into two clusters by PC1, while the loading plot (Fig. 8B) showed that most of the signals of quercetin and kaempferol derivatives, vitexin-2-*O*-rhamnoside and caffeic acid were located at the left side of PC1, corresponding to the position of absolute ethanol extracts obtained via sonication in the score plot. This revealed that the sonication was able to extract higher level of these phenolics than soaking.

PCA together with the PLS-DA results demonstrated that sonication combined with absolute ethanol was the most suitable extraction condition to produce *N. oleracea* extracts with high levels of valuable phenolics and hence high DPPH scavenging and α-glucosidase inhibitory activities. This effectiveness may be due to the combined effect of the ultrasound energy of sonication and the better ability of absolute ethanol to penetrate the cell wall of *N. oleracea*. Statistical analysis of the TPC, DPPH scavenging and α-glucosidase inhibitory activities of the different *N. oleracea* extracts further confirmed the effectiveness of sonication and absolute ethanol for producing extracts with high phenolic content and hence high bioactivities (Fig. 9).

CONCLUSIONS

PLS and RF regression models for the prediction of DPPH free radical scavenging and α-glucosidase inhibitory activities of *N. oleracea* were validated and compared. The RF model showed higher tendency of overfitting compared to PLS, especially in the prediction of DPPH free radical scavenging activity. Both the regression coefficient from PLS and the variable importance from RF revealed quercetin and kaempferol derivatives, caffeic acid and vitexin-2-*O*-rhamnoside to be significant contributors to the DPPH scavenging and α-glucosidase inhibitory activities of *N. oleracea*. Among the various extraction methods and ethanol ratios tested, sonication combined with absolute ethanol was able to extract the highest amount of these metabolites and hence contributed to highest DPPH scavenging and α-glucosidase inhibitory activities of this plant. This work reveals that the phenolics in *N. oleracea* are potent antioxidants and α-glucosidase inhibitors and that this plant has great potential for development of phenolic-rich food products.

ACKNOWLEDGEMENTS

This study was financially supported by Universiti Putra Malaysia under Research University Grant Scheme (RUGS) (9362700). The first author also gratefully acknowledges support from the Ministry of Higher Education Malaysia for the scholarship.

REFERENCES

1 Pandjaitan N, Howard LR, Morelock T and Gil MI, Antioxidant capacity and phenolic content of spinach as affected by genetics and maturation. *J Agric Food Chem* **53**:8618–8623 (2005).
2 Abdel-Farid IB, Hye KK, Young HC and Verpoorte R, Metabolic characterization of *Brassica rapa* leaves by NMR spectroscopy. *J Agric Food Chem* **55**:7936–7943 (2007).
3 Saleem M, Lupeol, a novel anti-inflammatory and anti-cancer dietary triterpene. *Cancer Lett* **285**:109–115 (2009).
4 Ayouni K, Berboucha-rahmani M, Kim HK, Atmani D, Verpoorte R and Choi YH, Metabolomic tool to identify antioxidant compounds of *Fraxinus angustifolia* leaf and stem bark extracts. *Ind Crop Prod* **88**:1–13 (2016).
5 Kim SH, Cho SK, Hyun SH, Park HE, Kim YS and Choi HK, Metabolic profiling and predicting the free radical scavenging activity of guava (*Psidium guajava* L.) leaves according to harvest time by ^{1}H-nuclear magnetic resonance spectroscopy. *Biosci Biotechnol Biochem* **75**:1090–1097 (2011).
6 Tahir HE, Xiaobo Z, Tinting S, Jiyong S and Mariod AA, Near-infrared (NIR) spectroscopy for rapid measurement of antioxidant properties and discrimination of sudanese honeys from different botanical origin. *Food Anal Methods* **9**:2631–2641 (2016).
7 Pariyani R, Ismail IS, Azam AA, Abas F and Shaari K, Identification of the compositional changes in *Orthosiphon stamineus* leaves triggered by different drying techniques using ^{1}H NMR metabolomics. *J Sci Food Agric*; https://doi.org/10.1002/jsfa.8288 (2017).
8 Paudel L, Wyzgoski FJ, Giusti MM, Johnson JL, Rinaldi PL, Scheerens JC *et al.*, NMR-based metabolomic investigation of bioactivity of chemical constituents in black raspberry (*Rubus occidentalis* L.) fruit extracts. *J Agric Food Chem* **62**:1989–1998 (2014).
9 Abdul-Hamid NA, Abas F, Ismail IS, Shaari K and Lajis NH, Influence of different drying treatments and extraction solvents on the metabolite profile and nitric oxide inhibitory activity of Ajwa dates. *J Food Sci* **80**:H2603–H2611 (2015).
10 Mediani A, Abas F, Khatib A, Tan CP, Ismail IS, Shaari K *et al.*, Phytochemical and biological features of *Phyllanthus niruri* and *Phyllanthus urinaria* harvested at different growth stages revealed by ^{1}H NMR-based metabolomics. *Ind Crops Prod* **77**:602–613 (2015).
11 Yuliana ND, Khatib A, Verpoorte R and Choi YH, Comprehensive extraction method integrated with NMR metabolomics: a new bioactivity screening method for plants, adenosine A1 receptor binding compounds in *Orthosiphon stamineus* Benth. *Anal Chem* **83**:6902–6096 (2011).
12 Breiman L. Random forests. *Mach Learn* **45**:5–32 (2001).
13 Beckmann M, Enot DP, Overy DP and Draper J, Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. *J Agric Food Chem* **55**:3444–3451 (2007).
14 Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C *et al.*, Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complement Altern Med* **2013**:298183 (2013).
15 Kalhan SC, Guo L, Edmison J, Dasarathy S, McCullough AJ, Hanson RW *et al.*, Plasma metabolomic profile in nonalcoholic fatty liver disease. *Metabolism* **60**:404–413 (2011).
16 Lanz C, Patterson AD, Slavík J, Krausz KW, Ledermann M, Gonzalez FJ *et al.*, Radiation metabolomics. 3. Biomarker discovery in the urine of gamma-irradiated rats using a simplified metabolomics protocol of gas chromatography–mass spectrometry combined with random forests machine learning algorithm. *Radiat Res* **172**:198–212 (2009).
17 Kovalishyn V, Aires-de-Sousa J, Ventura C, Elvas Leitão R and Martins F, QSAR modeling of antitubercular activity of diverse organic compounds. *Chemom Intell Lab Syst* **107**:69–74 (2011).
18 Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP and Feuston BP, Random Forest: a classification and regression tool for compound classification and QSAR Modeling. *J Chem Inf Comput Sci* **43**:1947–1958 (2003).

19 Acharjee A, Kloosterman B, de Vos RCH, Werij JS, Bachem CWB, Visser RGF *et al.*, Data integration and network reconstruction with ~omics data using random forest regression in potato. *Anal Chim Acta* **705**: 56–63 (2011).

20 Chanwitheesuk A, Teerawutgulrag A and Rakariyatham N, Screening of antioxidant activity and antioxidant compounds of some edible plants of Thailand. *Food Chem* **92**:491–497 (2005).

21 Lee SY, Mediani A, Nur Ashikin AH, Azliana ABS and Abas F, Antioxidant and α-glucosidase inhibitory activities of the leaf and stem of selected traditional medicinal plants. *Int Food Res J* **21**:165–72 (2014).

22 Lee SY, Abas F, Khatib A, Ismail IS, Shaari K and Zawawi N, Metabolite profiling of *Neptunia oleracea* and correlation with antioxidant and α-glucosidase inhibitory activities using ^{1}H NMR-based metabolomics. *Phytochem Lett* **16**:23–33 (2016).

23 Javadi N, Abas F, Hamid AA, Simoh S, Shaari K, Ismail IS *et al.*, GC–MS-based metabolite profiling of *Cosmos caudatus* leaves possessing alpha-glucosidase inhibitory activity. *J Food Sci* **79**:1130–1136 (2014).

24 Jaime L, Vázquez E, Fornari T, López-Hazas M del C, García-Risco MR, Santoyo S *et al.*, Extraction of functional ingredients from spinach (*Spinacia oleracea* L.) using liquid solvent and supercritical CO_2 extraction. *J Sci Food Agric* **95**:722–729 (2015).

25 Mediani A, Abas F, Khatib A, Maulidiani H, Shaari K, Choi YH *et al.*, ^{1}H NMR-based metabolomics approach to understanding the drying effects on the phytochemicals in *Cosmos caudatus*. *Food Res Int* **49**:763–770 (2012).

26 Liaw A and Wiener M, Classification and regression by randomForest. *R News* **2**:18–22 (2002).

27 Oshiro TM, Perez PS and Baranauskas JA, How many trees in a random forest?, in *Machine Learning and Data Mining in Pattern Recognition: Lecture Notes in Computer Science*, Vol. 7376, ed. by Perner P. Springer, Berlin, pp. 154–168 (2012).

28 Maulidiani, Abas F, Khatib A, Shitan M, Shaari K and Lajis NH, Comparison of partial least squares and artificial neural network for the prediction of antioxidant activity in extract of Pegaga (*Centella*) varieties from ^{1}H nuclear magnetic resonance spectroscopy. *Food Res Int* **54**:852–860 (2013).

29 Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML *et al.*, A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding? *Anal Chim Acta* **879**:10–23 (2015).

30 Hung HY, Qian K, Morris-Natschke SL, Hsu CS and Lee KH, Recent discovery of plant-derived anti-diabetic natural products. *Nat Prod Rep* **29**:580–606 (2012).

31 Jo SH, Ka EH, Lee HS, Apostolidis E, Jang HD and Kwon YI, Comparison of antioxidant potential and rat intestinal α-glucosidases inhibitory activities of quercetin, rutin, and isoquercetin. *Int J Appl Res Nat Prod* **2**:52–60 (2009).

32 Li YQ, Zhou FC, Gao F, Bian JS and Shan F, Comparative evaluation of quercetin, isoquercetin and rutin as inhibitors of α-glucosidase. *J Agric Food Chem* **57**:11463–11468 (2009).

33 Li H, Song F, Xing J, Tsao R, Liu Z and Liu S, Screening and structural characterization of α-glucosidase inhibitors from hawthorn leaf flavonoids extract by ultrafiltration LC-DAD-MSn and SORI-CID FTICR MS. *J Am Soc Mass Spectrom* **20**:1496–1503 (2009).

34 Oboh G, Ademiluyi AO, Akinyemi AJ, Henle T, Saliu JA and Schwarzenbolz U, Inhibitory effect of polyphenol-rich extracts of jute leaf (*Corchorus olitorius*) on key enzyme linked to type 2 diabetes (α-amylase and α-glucosidase) and hypertension (angiotensin I converting) in vitro. *J Funct Foods* **4**:450–458 (2012).

35 Sato Y, Itagaki S, Kurokawa T, Ogura J, Kobayashi M, Hirano T *et al.*, In vitro and in vivo antioxidant properties of chlorogenic acid and caffeic acid. *Int J Pharm* **403**:136–138 (2011).

36 Kek SP, Chin NL, Tan SW, Yusof YA and Chua LS, Classification of honey from its bee origin via chemical profiles and mineral content. *Food Anal Methods* **10**:1–12 (2016).